REPUBLIC OF TURKEY

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electricity And Computer Engineering

DEFINING LOCAL RISK POINTS AND DISTANCE VARIABLES AND STUDYING EFFECTS ON POLICY RENEWAL SCORE DEVELOPED WITH GENERAL LINEAR ALGORITHM

Vedat GÜNEŞ

Master Thesis

Supervisor

Asst. Prof. Dr. Oğuz ATA

Istanbul, 2022

DEFINING LOCAL RISK POINTS AND DISTANCE VARIABLES AND STUDYING EFFECTS ON POLICY RENEWAL SCORE DEVELOPED WITH GENERAL LINEAR ALGORITHM

Vedat GÜNEŞ

Electricity and Computer Engineering Department

Master Thesis

ALTINBAŞ UNIVERSITY

2022

The thesis titled DEFINING LOCAL RISK POINTS AND DISTANCE VARIABLES AND STUDYING EFFECTS ON POLICY RENEWAL SCORE DEVELOPED WITH GENERAL LINEAR ALGORITHM prepared by VEDAT GÜNEŞ and submitted on 12/08/2022 has been **accepted unanimously** for the degree of Master of Science in Electrical and Computer Engineering.

Asst. Prof. Dr. Oğuz ATA

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Oğuz ATA	Faculty of Software Engineering, Altınbaş University	_____
Asst. Prof. Dr. Doğu Çağdaş ATİLLA	Faculty of Electrical-Electronics Engineering, Altınbaş University	_____
Asst. Prof. Dr. Aytuğ BOYACI	Faculty of Electrical-Electronics Engineering, Altınbaş University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

Submission date of the thesis to Institute of Graduate Studies: __ /__ / __

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Vedat GÜNEŞ

DEDICATION

I dedicate this study to my dear wife and my sweet daughters. They always support me to have success in my study and career.

PREFACE

The work described in this thesis/dissertation has been supported by Anadolu Anonim Türk Sigorta Şirketi, called Anadolu Sigorta. Anadolu Sigorta supplied the data set that we examined while studying this thesis. I gratefully thank Anadolu Sigorta for their collaboration and support.

ABSTRACT

DEFINING LOCAL RISK POINTS AND DISTANCE VARIABLES AND STUDYING EFFECTS ON POLICY RENEWAL SCORE DEVELOPED WITH GENERAL LINEAR ALGORITHM

Güneş, Vedat

M.Sc., Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Dr. Oğuz ATA

Date: August/2022

Pages: 72

Insurance covers customer risks. The motor insurance domain is the biggest and most well-known topic in the insurance sector by customers. Claim management is the most crucial process for the customers. They get coverage payments due to claim flow defined by the insurers. Insurers always try to decrease claims. To manage claims, the most significant expenditure for insurers, insurers develop pricing models, risk assessments, and fraud management flows. Location is one of the most critical parameters for the analytics insurers develop.

Motor insurance covers vehicle risks for insurance customers. Vehicle risk management mainly uses location and location-based analytics. The places and route of the vehicle help insurers define the policy's risk level. Almost every model gets this input to consider the location of the vehicle. The location is also vital because activities that happened at that location are also significant to consider and affect the calculation of the analytical models.

Accident reports record the location of the vehicle, and this parameter could be used as a parameter to define locations' risk levels more accurately. Based on the distance between the risky location and the customer, we can identify a new approach for the insurers to consider and define more accurate models to manage the claim and risk level of the customers.

Keywords: Distance to Risky Locations, Forecasting Customer Renewal, Insurance Claim, Location Analytics, Location Base Risk

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
ML	:	Machine Learning
SBM	:	Sigorta Bilgi Merkezi
IBNR	:	Issued But Not Reported
DL	:	Deep Learnign
GML	:	General Linear Model
GLMMs	:	Generalized linear mixed models
GNMs	:	Generalized Non-Linear Models
NPS	:	Net Promoted Score
CRISP-DM	:	Cross Industry Standard Protocol for Data Mining
KNIME	:	Konstanz Information Miner
CLI	:	Command Line Interface
GUI	:	Graphical User Interface
API	:	Application Programmable Interface
EDA	:	Exploratory Data Analysis
SVM	:	support vector machine
PCA	:	Principal Component Analysis
SVD	:	Singular Value Decomposition
GIS	:	Geographical Information System

LIST OF SYMBOLS

μ : Mean

$\in$: The error value of linear model

β : Linear predictor

g : Link function

1. INTRODUCTION

Insurance is one of the oldest sectors in the finance industry. People purchase insurance policies to minimize risk on the assets they own. Insurance policies cover every kind of asset and cover customers based on the terms included in the insurance policy. Car, residential, and health insurance types are essential topics in the insurance market in Turkey.

Preparing policies is the first step and the beginning of the communication with the customers and insurance companies. Underwriting and claim processes are the main flows in the insurance life-cycle. Risk assessment of a policy is the key feature of the insurance income model.

This study will discuss car insurance risk calculation and price optimization of the policy by accident reports. Insurance Information and Monitoring Center (Sigorta Bilgi Merkezi - SBM) collects accident data and shares the data with insurance companies. Anadolu Sigorta's and SBM accident reports will be included in the risk calculation and policy premium optimization. Firstly, based on the accidents that happen in coordination, we calculate the risk level of that point. Then, the calculated feature will be included in the price premium optimization to study the effect of the changes. The insurance company can decide to include the feature in its risk formula or offer an optimized premium to the policyholder.

The study covers only personal car insurance risk calculation and price optimization. Other car types were excluded from the data set.

1.1 BACKGROUND

Insurance is one of the essential financial domains that sold products by insurers to safeguard the policyholders and their property against the risk of loss [1]. The customer buys a policy defined the coverage by the insurance customer makes regular payments, which are called premiums, to the insurer. If the customer has an issue that is covered by the insurance policy called a claim, the insurer will pay that claim for the loss covered under the policy that the customer buys. If there is no claim-defined timeline for the policy, the insurer does not pay anything back to the customer. Instead, the money is pooled by the insurer. The insurer pays all the claims under the policy coverage from the pool collected from all customers who have policies with the same insurer. There are five main steps in the insurance business those insurers should manage to generate profitable revenue [2],

I. The Quote Management

II. Risk Assessment

III. Monitoring

IV. Claim Management

V. Renewing

Insurers have actuary departments to process data and create models to manage the main five processes [3]. The core action managed by the actuary department is a risk. They create models to estimate the risk values of their customers, entities, and all the objects subject to the insurance policy [4]. Many variables are used to calculate risks, premiums, and claims [5], [6].

Insurance has different focus areas mainly divided into two; motor insurance and mon-motor insurance domains [7]. Motor insurance covers vehicle-focused risks. The driver must have an automobile insurance policy before driving the vehicle [8]. The insurance policy has coverages that protect the customer against liability when an accident occurs. Plus, the vehicle and the stakeholders in the accident are in the scope of the policy coverage. It provides financial compensation to cover injuries caused to people or their property. Non-motor insurance covers all the risks and focuses other than a vehicle, such as health, personal accidents, travel, fire, burglary, marine, engineering, etc. [9].

The location of the customer, vehicle or action related to the policy coverage is an essential variable used in the actuarial models. The motor insurance domain uses the location variable while calculating the risk score of the entity (vehicle or customer). The risk score affects the quote and premium amount of the policy. If the risk score is minor, customer premium is defined closer to the base premium or vice versa.

Claim history is significant for the renewal process. When the customer wants to renew the policy that was held for the last year, the models check the claim history. Higher the claim amount, the customer gets a higher quote. Not only is the amount of the claim essential for calculating the renewal quote, but the event's location is critical. Location is included in the actuarial models to calculate the premium amount [10].

The location variable is covered at the city level in the actuarial models. Customer and the vehicle city are essential for the models. In detail, the city is wider to manage the correct risk value. Plus,

when the entire city risk is determined, micro risk levels can not be included, and the entire city risk is shared with all customers who live in the city. County, road, and street-level risk calculations correctly handle the risk distribution for the customers who are not risky.

Accident reports have location information. This study will focus on accident locations and create new variables for actuarial and related analytical models. Risky points will be located on the map, and then the distance will be calculated to study the effects on the renewal score estimation model.

1.2 OBJECTIVES

Insurers try to manage customer portfolios by estimating their risk. Risk estimation is a critical factor in profitability. Actuarial models were developed to identify the risk values of the customer portfolio. Insurers have potentially high-risk means they have a substantial amount of claims.

Claim management is a critical process for profitability for insurers. A claim results from correct customer selection, risk analysis and accurate premium optimization models. As a result, the claim is not the reason; it is a result.

Motor insurance risk calculation models use various variables related to customer and vehicle. This study aims to create two new variables; risky points and customer distance to the nearest risky points;

i *Risky Locations:* Accident reports will be used to define the risky points, and the result will be segmented into four main groups.

ii *Distance:* Customer address information will be used to measure the distance to the nearest risky point.

Insurers try to keep their existing customers instead of gaining new customers. It is more expensive to gain a new customer in every channel. Anadolu Sigorta uses a renewal estimation model to focus on the correct customer set to increase the number of renewed policies. The ultimate goal of the process is to increase insurer profitability by increasing revenue. Risky locations and distance will be the new input variables for the renewal score estimation model. *The renewal estimation model* is a data analytics flow that uses machine learning algorithms to estimate the target variable. As a result of the study, we will share how the new variables affect the result calculated before adding them to the flow.

1.3 THESIS LAYOUT

This thesis will cover the topics in the chapters defined below:

Chapter 2 performs a literature review of actuarial risk calculation and location base analytics used in the insurance domain.

Chapter 3 describes the datasets, methods, and software used in this study. We share theocratical background on proposed methodologies.

Chapter 4 gives the calculation process of the new variables and adds them to the analytical model. Then the results will be compared between the initial analytical and update models.

Chapter 5 shows the conclusion of the study and proposes further improvements.

2. LITERATURE REVIEW

In this section, we discuss existing actuarial models, which are the data processing activities for insurance. Actuary sciences have started to process data since the beginning of the insurance topic. Based on the collected data about the customer or subjects, insurance models were developed to identify the risk. Risk score impacts quotes, premiums, and claim management.

Actuarial vocation has started to grow and process more data than ever. Statistics and traditional methods do not help actuary to handle all the risks subject to the insurance sector [11]. Big data analytics is now an essential topic for insurers. There are essential developments in Artificial Intelligence (AI) on using data effectively, both structured and unstructured. Machine Learning (ML), the subset of AI, has numerous algorithms to create value by understanding the inner bonds of variables and making predictions on data. The specialized algorithms collected under the Deep Learning (DL) collection focused on computer vision, Neuro-Linguistic Programming, and image and video processing. Figure 2.1 shows us the relationship between the three concepts; artificial intelligence, machine learning, and deep learning [12].

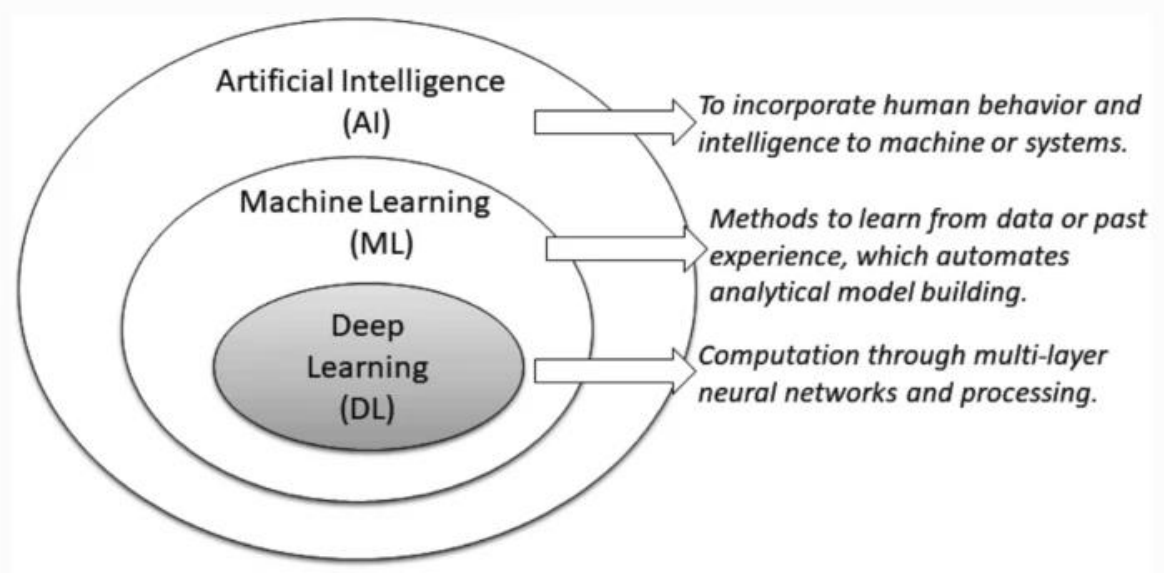

Figure 2.1 Relation Between AI, ML, and DL and Short Descriptions [12]

2.1 INSURANCE DOMAINS

2.1.1 Non-Life Insurance Domains

Insurance has various types of products. The collection of insurance products other than life is called non-life. From the perspective of an actuary non-life domain have two main functions:

i *Pricing Actuary:* Designs products and defines the pricing strategy of the insurance products.

ii *Claim Prediction Actuary:* Predicts the insurance claims for the whole portfolio of the insurers called cash flow prediction of claims. These activities are subject to define accounting, product development, and risk management.

The main three sub-domain of non-life insurance products are motor, non-motor, and health insurance.

2.1.2 Life Insurance Domain

The insurance domain related to life subjects is called Life and Pension Insurance. This domain offers their customer protection against disability and death. Life insurers try to predict the mortality of their customers and make multi-year contracts with them. Insurers collect payments from the customers and make investments to increase the amount. Thus they build good financial models to manage their income because they need to manage the uncertainties of their customers' mortality guarantee [3].

2.1.3 Risk Management and Changes in Trends Over Data and Technology

Reinsurance is selling the risk and distributing it into pieces. In other words, ensuring the insured issue again to a new insurer [13]. Insurers take risks by selling their products to their customers. They collect risk and need to deliver or sell the risk to protect themselves from paying many claims.

Technology and climate change are affecting many sectors, as well as insurance. Technical damages cause huge claims like cyber attacks, data leakages, and privacy violations. These were operations issues before, but they are getting more damageable than ever with technological transformation.

Climate changes affect almost everything. Insurers try to manage impactful damages and claims caused by environmental issues.

Insurance is changing like many other domains with technological developments. There are either opportunities or difficulties. Insurers develop new products to increase their revenue. They sign new reinsurance agreements to decrease claims and risk. In the decade, both topics are evolving to a new level [14].

2.2 ACTUARIAL MODELS

Data have created a new ecosystem since the 2000s [15] [16]. The data era has started with the internet and fastened up with mobile devices. Everybody has become a data creator by using mobile devices; texting, taking photos, sharing likes, commenting on everytinh etc. The massive amount of data created allows data scientists to develop various models to extract insights about the customers in every sector. Insurance is the sector that imposed this wave very late because the insurers have an actuary department that processed data to develop models for them [17].

Impactful advantages in Artificial Intelligence (AI) and Machine Learning (ML) have started to define actuarial models from scratch. AI and ML models bring capabilities to change existing insurance models and create new models. The main game changer is to have different types and a vast amount of data [Data amount in ML and DL]. Deep learning (DL), the subset of ML, also has impactful capabilities to identify insights by using hidden layers and connections in data [18], [19].

Actuarial problems are mainly addressed to be solved with a regression model. Even if it seems not to be a regression model at first determination, actuaries transfer the issue to be solved with a regression equation. The most used actuarial models are exampled in table 2.1 and summarized in the list below;

1. Insurance policies mainly can be for short terms. Customers want to renew the policies at the end of the term. Insurance policies are priced with GLM models of frequency (The primary method is Poisson rate regression) and severity (gamma regression).
2. The chain-ladder calculates claims reserve, which can be described as a cross-classified log-linear regression according to the GLM formulation [20].
3. The chain-ladder analysis estimates the reserver claim amount for the next year by analyzing the occurred claim amounts [21]. The model built in 1993 uses different linear regression models to predict the coefficients and the factor value resulting by using a single weighted linear regression algorithm [22].
4. Incurred But Not Reported (IBNR) defines the claim amount, which is estimated, but the exact value amount has not been decided, only estimated. Generalized linear mixed models (GLMMs) and Bayesian hierarchical modes are used to estimate IBNRs [22].
5. Mortality tables and expected experience are life insurance subjects that can be modeled as a Poisson regression problem. The model uses the expected mortality rate as input and creates the number of deaths figure with an offset taken as the essential variable for the risk [23]
6. The insurers use regression models to determine a life-table-based mortality experience of their portfolio, which gets their customers' gender, ages, and other essential variables and estimates the number of deaths as output [24].
7. Life insurance uses mortality tables, and models are developed as regression problems. GLM is used to estimate the accurate results for the mortality problem in life insurance [25]. Plus, some problems are modeled and solved using Generalized Non-Linear Models (GNMs) [26].

8. The demographic techniques known as the Near Extinct Generations methods [27], which are used to derive mortality rates at older ages from death counts, can be solved as a regression problem within the scope of the GLM framework [28].
9. Life policies have cash-flow projections to define income and profitability balance—the cash-flow projections generated by valuation models in life insurance domains. Valuation models use a group of life insurance policies to estimate the future cash flows, which helps to have reserves or embedded value. Life valuation models take a long to get the result, so using regression models helps calibrate them [29]
10. Nested stochastic simulations are a regression problem which is a problem to avoid in the life insurance domain. This concern needed to be handled in Solvency II or the Swiss Solvency Test by calibrating the regression model, which is an excellent way to solve the problem [30]

Table 2.1 The Liste Of The Problems In Actuarial Science Solved By Regression

Number	Description	Feature Vector	Output
1	Short-term pricing	Policyholder details, such as age, gender, marital status, credit score, postal code of residence, and details of the insured item, such as, on a motor policy, brand, age, power of the engine	Frequency/Severity of Claim, or Pure Premium (Tweedie GLM)
2	Over-dispersed Poisson IBNR model	Accident year, reporting year	Incremental claim amount
3	Mack chain-ladder model	Cumulative claims amount at time t-1	Cumulative claims amount at time t
4	Hierarchal IBNR models	Refer to 2&3	Cumulative claims amount at time t
5	AvE analysis of mortality	Expected Number of deaths	Coefficient of the regression model
6	Portfolio specific mortality model	Age, gender, portfolio characteristics	Mortality rate
7	Mortality forecasting	Year, age	Mortality rate
8	Old-age mortality estimation	Year of birth, age at deaths	Number of deaths
9	Life valuation approximation	Age, gender, portfolio characteristics	Reserve value
10	Life valuation approximation	Value of assets (bonds, equities, options, swaptions) in scenario i	Reserve value

As it is seen in Table 2.1, insurance business run on data and data processing. The new era presents new advantages to processing data using AL, ML, and DL, so insurance has impactful advantages.

2.3 RENEWAL MODEL FOR MOTOR INSURANCE

A renewal prediction algorithm is a model that predicts how the customer is likely to renew the policy. Since it will help us focus on customers we know are not likely to renew. Insurers, like any other companies, tend to save their customers because it is cheaper than gaining new customers who do not have a relationship before. Insurers focus on customers with low renewal scores and define actions to keep them. They also are in touch with the customers who renew their policies.

Renewal models in insurance have different types of calculations. There are models built to examine and estimate the score more accurately. One notable and essential renewal model type is the uncertain renewal model [31]. In uncertainity, age replacement policy was studied, and its results applied to the renewal process [32].

The renewal process proceeded to be analyzed by applying calculus to the developed algorithms. The new research proposed to study the uncertain calculus of the integration and differentiation of the uncertain renewal process [33]. The renewal process was studied differently by adding off-time and on-time variables to the modeling flow to alternate the renewal problem [34].

Renewal models studied in insurance as a complicated problem. A linear model was used to examine the problem, and a generalized linear model was tested for better results with different variables. In this thesis, we study to add two new variables to test the renewal score and try to study the results.

2.4 LOCATION ANALYTICS IN INSURANCE

Location defines where the event happens in insurance.[35] The location parameter is vital for catastrophic claims to identify how much the disease affects the customer base. Insurers have map base analysis and use geographical information systems to identify the effect of the claim. It is now imperative for the motor insurance domain to use location variables. Claim experts handle the claim file and investigate the accident based on location. Insurers define vehicle services as towing or on-place repair. They need customer location information to reach them correctly and quickly [36].

Insurers collect location data mainly by using telematics [37]. It is also possible to have the route information via telematics. Unfortunately, it is not mandatory to use telematics in every country. Insurers developed mobile applications to serve information and services digitally to their customers [38]. Mobile applications are used to collect data via cellular phones like; competitor applications installed, location, customer complaints etc. [39].

Insurers develop location-based value-added services for their customers to increase net promoted score (NPS) [40]. The critical data variable is the location to create, develop, and serve these services.

2.5 EUCLIDEAN DISTANCE

The distance approach is widely used. Euclidean squared distance helps us to calculate the distance between two different coordinates. In this study, we will locate risky points where the maximum number of accidents happened and pick the smallest Euclidean Squared Distance that gives us the result in this approach [41]. Various techniques have evolved from the fundamental approach, such

as Pearson's correlation, distance correlation, and angular distance. [42], [43]. Equation 2.1 shows the formula of Euclidian squared distance.

$$SSD(Sum\ of\ Squared\ Distance) = \sum\_(t = 1)\^N \square (P\_t\^1 - P\_t\^2\ )\^2 \quad (2.1)$$

The Euclidean distance gives us the point-to-point distance between two coordinates. Minimizing the sum of the squared difference between two coordinates is not the path distance in two locations. To find the exact path distance, we need to have route information. Due to the lack of data and data privacy concerns [44] of Turkish customers, it is not possible to have route information and exact path to calculate the accurate distance between customer address and risky location.

Figure 2.2 The actual road distance vs. the Euclidean distance [45]

3. METHODOLOGY

3.1 DATASET

There are two types of accidents tracked in insurance industry;

- *Bodily insured accidents*
- *Property damage accidents*

In this study, we will process only Property Damage Accidents. The insurance Information and Identification Center (SBM) collects all the accident reports and analyzes by using them. Figure 3.1 shows the distribution of the accidents that happened beginning from 2008 till 2021.

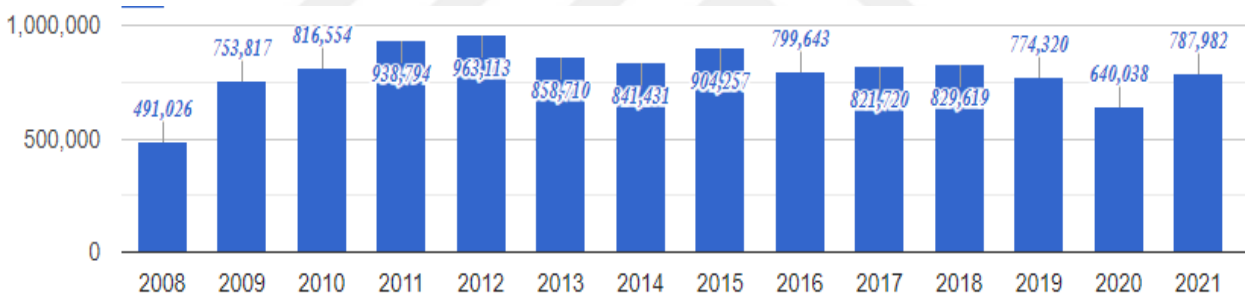

Figure 3.1 Number of Accident Reports per Year 2008 - 2021

In this study, we will fallow the flow defined in the figure 3.2;

- Prepare a data set to process accident reports.
- Create segments on the accident reports based on the number of accidents that occurred in the same location
- Visualize the segmented locations and accident information on a map visual
- Get the policy list and address details of the customers by joining the claim no and policy no

- Extract the longitude and latitude values from Google API
- Calculate the distance between the customer address location and risky points calculated based on the accidents
- Assign the nearest risky point to the customer's address
- Use the distance variable in the renewal score calculation developed by using General Linear Model for motor insurance policies.

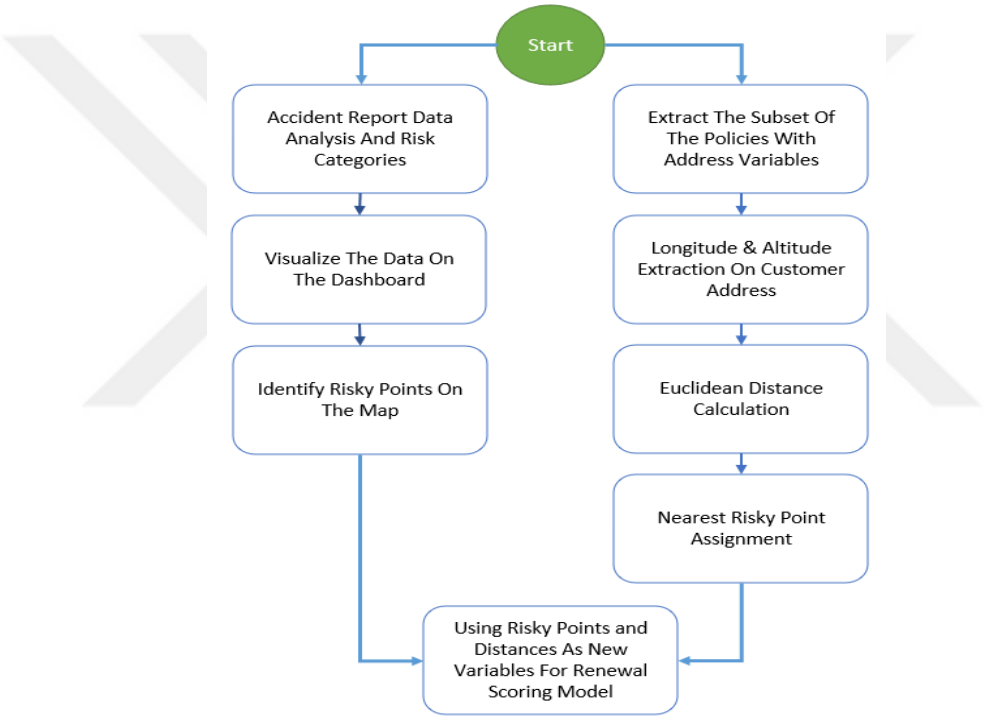

Figure 3.2 Data Preparation And Analysis Flow Of The Study

The accident report data model has a common approach in Insurance Industry in Turkey. Insurers more or less store the accident report data in the same approach. Figure 3.3 shows the reference model for managing the data in databases. Insurers store the data and use it in their analysis, reporting and machine learning models to create insights to identify driver risks.

Accident reports have mandatory fields. The most complete and high-quality data is stored in these mandatory fields. Other variables have fewer quality data which is needed to be curated.

kttAccident table shown in the model keeps the location variables of the accident. We will focus on this information to identify the risk of the accident locations. *The longitude* and *latitude* of the accident address give us the exact point and help us to locate the accident on a map.

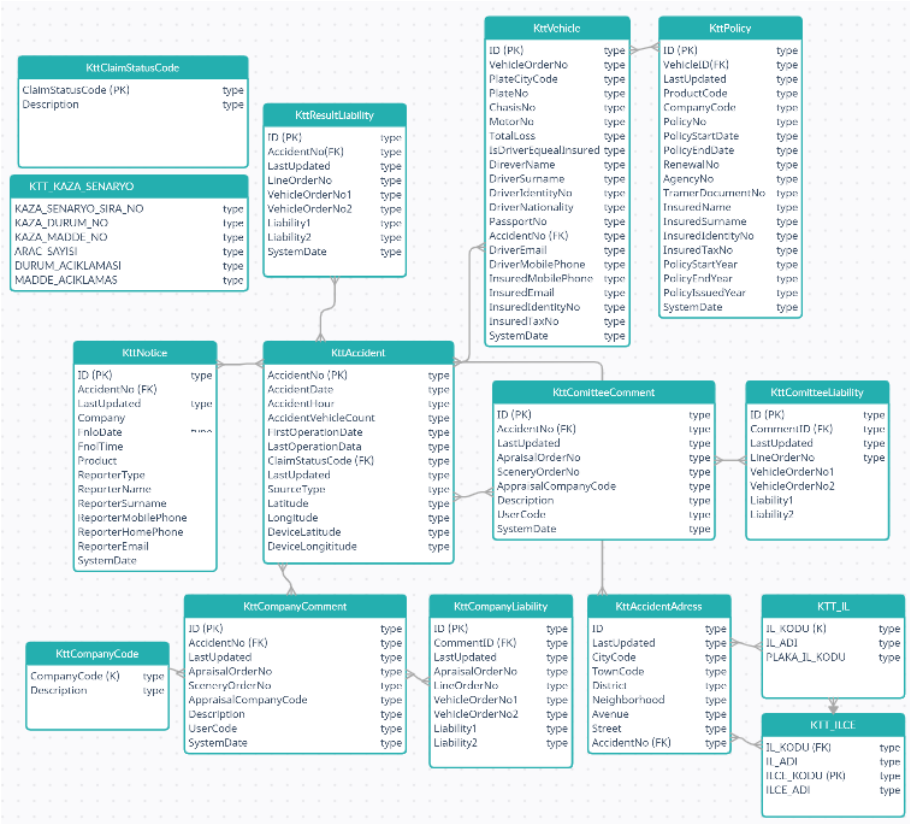

Figure 3.3 Accident Report Data Model

As is seen in the data model, there are many tables and variables. The data in the model shared in figure 3.3 are used in different business flows. Mainly this data is subject to the claim process of the insurers. Insurers share this data with the security forces to inform them about the accident. There are reports created to inform government units when they request.

In the thesis, we need to extract the coordinates of the accident. kttAccident table has location variables that define the exact point of the accidents. There is no location information in the table for each line. We excluded the lines which do not have coordinate values.

To prepare the data, we denormalize the data model into a single view. The variables added to the denormalized structure are shared in Table 3.1. This data set will be de primary data set for the first part of the study, segmenting the data and identifying the risky locations.

Table 3.1 Denormalized Data Variables

Variable Name	Variable Description
Company Code	There are more than 40 insurers in Turkey. This variable refers which the insurer has a policy with the accident stakeholder.
Accident Vehicle Count	In an accident, the number of the vehicle is important to verify the scope of the accident. This variable keeps the total number of vehicles taking place in the accident.
Accident No	Refers to the unique id in the accident detail tables.
Accident Date	The date when the accident happens
Vehicle Order No	The policy owner's vehicle order no in the occurred accident.
Vehicle Order No1	The first vehicle in the accident
Vehicle Order No2	The second vehicle in the accident
Liability1	Insurers pay claims based on the share of their customers in the accident. The share is distributed based on the liability of the customer. The first stakeholder liability.
Liability2	Second stakeholder liability.
Product Code	Accident report stakeholders can have different products like traffic or auto insurance in Turkey. Product code distinguishes the type of the insurance type.
City ID	In which city the accident happens
County ID	In which county the accident happens
Plate City Code	The city that the plate registered.
Plate No	Unique identifier of the car in the country
Chasis No	There is a unique number for every car called chassis no
Policy No	Customers purchase a policy from the insurer, and every customer has a unique policy no
Claim No	The insurer creates a claim file for the customer who has an accident. Claim no is the unique identifier for the claim request of the customer.
Renewal No	If the customer renews, his insurers do not prepare a new policy. Instead, they prepare an attachment to the main policy. Renewal No defines how many times the customer renews the policy at the same insurer.
Insured Identity No	The unique identifier for the customer.
Longitude	Y coordinates of the accident
Latitude	X coordinates of the accident

Accident reports are created for the accidents that happen in a location. For each accident, there is one accident report. Each stakeholder is defined in the accident report called *Accident Vehicle*. Accidents in the exact location are counted and assigned to the location. The value of the number defines the risk segment of the location. The study created four segments: small*, medium, high, and urgent*. All the accident reports counted for this analysis between 2011 and 2021.

We will use the variables listed in the table 3.2.

Table 3.2 Variables Used in the Risk Segmentation Data Set

Variable Name	Variable Description
Product Code	Insurer Company
Company Code	Accident Details Reference ID
Accident Date	When the accident happens
Number Of Accident	Insurance Product Code
Accident Latitude	X Coordinates of the Accident
Accident Longitude	Y Coordinates of the Accident
Risk Segment	Vehicle Plate No

The intervals of the numbers for the risk segment variable are listed below;

1- Small Risk 0 < Number of Accidents <= 250

2- Medium Risk 250 < Number of Accidents <= 500

3- High Risk 500 < Number of Accidents <= 1000

4- Urgent Risk 1000 < Number of Accidents

According to the distribution, the number of accidents addressed to the segments is shown in table 3.3. The bar distribution can be found in Appendix A.1.

Table 3.3 Accident Distribution According To The Risk Segmnets

Risk Segment	Total Number Of Accidents
Small Risk	623.225
Meidum Risk	39.480
High Risk	39.847
Urgent Risk	95.019

We will continue to study urgent points. Urgent points have an enormous number of accidents, affecting insurers more than others means the urgent points cause recurring claims. The list of all urgent points is shared in Appendix A1.1. İstanbul has the maximum number of urgent points, so we will continue to analyze İstanbul's data for this study.

Figure 3.4 shows the distribution of the urgent points (number of accidents greater than 1000). We developed a map visualization to show the locations in a map supplies as the coordinates of those points. We can extract the latitude and longitude values of the urgent locations in İstanbul. Table 3.4 lists the urgent points and their latitude and longitude values. All the urgent point lists for Türkiye can be found in Appendix A.1.

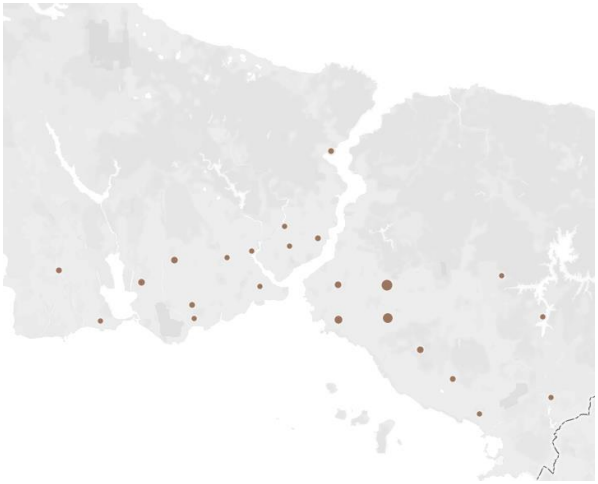

Figure 3.4 Urgent Point Visual of İstanbul

Table 3.4 Urgent Locations' Latitudes And Longitudes

Istanbul Urgent Points	Latitude	Longitude
EU1	41.03431	28.66148
EU2	40.98019	28.72067
EU3	41.02123	28.77809
EU4	41.04564	28.82472
EU5	40.9975	28.85055
EU6	40.98319	28.8536
EU7	41.04815	28.90045
EU8	41.055	28.93444
EU9	41.01686	28.94704
EU10	41.0814	28.98197
EU11	41.06006	28.98811
EU12	41.16229	29.04742
EU13	41.06861	29.02853
AS1	41.01894	29.05762
AS2	40.98187	29.05762
AS3	41.01823	29.12743
AS4	40.98333	29.12777
AS5	40.9498	29.17395
AS6	40.9184	29.22045
AS7	40.88094	29.25774
AS8	41.0287	29.29018
AS9	40.98479	29.34826
AS10	40.89823	29.35987

The second data set is customer and policy data. There is a link between accident reports and customer data set via *Claim No*. We use Claim No to get the related policy list. Policy data helps us prepare the customer and address list, which relate to the urgent risk locations.

Customers do not prefer to share their exact locations. Insurers try to collect location information by giving discounts and developing mobile applications to get permission to store location data via mobile phones. We use Google Geolocation API (Described in Section 3.3.1), *Place or pin a Marker on the map* service to get the coordinate values of the address. In order to get the location

values, we need the customer's city, county, district and road information. Using policy no data, we can get the customer address variables needed to extract location values.

Figure 3.5 has a sample list of variables listed to extract coordinates by using Google API. If the data does not have a missing parameter value, the code aggregates the location one level up. For example, if C1 has no road information, the location is extracted according to the district. Suppose there are no district information location values extracted according to county information. At every step of aggregation, the midpoint of the county, district, or road is determined for the coordinate extraction. No line does not have any county value. So at least we have the central location of the customer's county.

Policy No	Claim No	Renewal Status	Customer No	City	County	Distrcit	Road
P1	C1	1	Cus1	İSTANBUL	ÜSKÜDAR	KISIKLI	Kısıklı
P2	C2	0	Cus2	İSTANBUL	PENDİK	GÜZELYALI	İstiklal
P3	C3	1	Cus3	İSTANBUL	KADIKÖY	MECLİS	
P4	C4	1	Cus4	İSTANBUL	KADIKÖY	ERENKÖY	Ethem Efendi
P5	C5	1	Cus5	İSTANBUL	ŞİŞLİ	ESENTEPE	Kre Şehitleri
P6	C6	1	Cus6	İSTANBUL	KÜÇÜKÇEKMECE	KARABURNA-YENİ	
P7	C7	1	Cus7	İSTANBUL	KÜÇÜKÇEKMECE	HALKALI MERKEZ	Fatih Caddesi
P8	C8	1	Cus8	İSTANBUL	TUZLA	ÇARŞI	Sahil
P9	C9	0	Cus9	İSTANBUL	ÜMRANİYE	ÇAMLIK	İkbal Caddesi
P10	C10	0	Cus10	İSTANBUL	ÜMRANİYE	ÇAMLIK	İkbal Caddesi

Figure 3.5 Sample Policy List with Customer Address Variables

The data set used in this study will be added to the thesis attachments by applying data privacy regulation rules.

For the next step, we have location information for each line in the data set. The next step is calculating the distance between urgent segmented risk locations. We use the Euclidean distance method to calculate the distances.

Euclidian distance formula was applied to the dataset using the python math library to get the distance. The definition of the formula used to calculate the distance between two points is defined in Equation 1. [46] [47];

(x1 1, y1 1) and (x2 2, y2 2) is

d = √[(x2– x1)2 + (y2 – y1)2].

Equation 2- Euclidean Distance Formula

$$SSD(Sum\ of\ Squared\ Distance) = \sum\_(t = 1)\text{^}N \text{▒}(P\_t\text{^}1 - P\_t\text{^}2\ )\text{^}2 \qquad (3.3)$$

Figure 3.6 shows how the points are defined in the grid. Also, the Euclidean distance equation was added below;

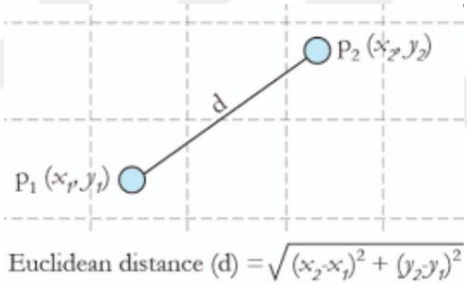

Figure 3.6 Euclidean Distance Calculation

The distance between two coordinates is calculated by using the python math library. Euclidian distance calculation used for measuring the distance as defined below;

DistanceY = Y2 - Y1

DistanceX = X2 - X1

a = math.pow(math.sin(DistanceX / 2), 2) +

math.cos(X1) *math.cos(X2) *

math.pow(math.sin(DistanceY / 2), 2)

c = 2 * math.asin(math.sqrt(a))

To calculate the distances between the customer address and urgent level risky points, we use python language (Described in section 3.11) and integrated it into KNIME Analytics Platform (Described in section 3.3.3) as a flow in figure 3.7.

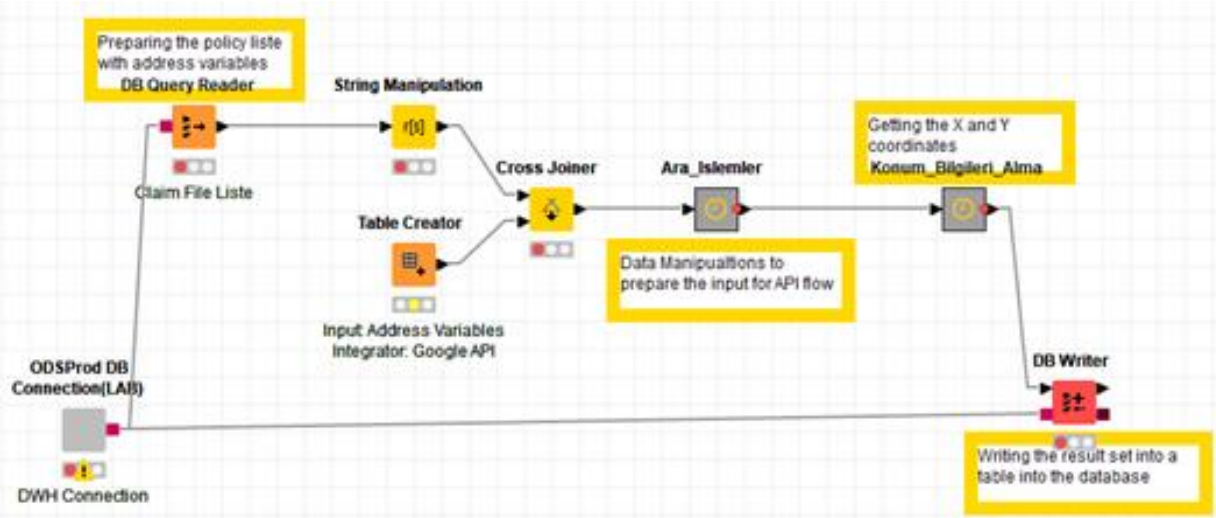

Figure 3.7 KNIME Flow For Coordinate Identification And Distance Calculation

The list of customer addresses list is the input of the flow, and we get the distance values to each risky point given to the flow. The flow identifies the nearest risky point to the customer location based on the calculated Euclidian distances. The final view of the data is shown in Figure 3.8.

Policy No	Claim No	Renewal Status	Csutomer ID	City	County	District	Urgent Point	Distance
P1	C1	1	Cus1	İSTANBUL	BAHÇELİEVLER		Anadolu_2	6.973593088
P2	C2	0	Cus2	İSTANBUL	BEYLİKDÜZÜ		Anadolu_2	6.973593088
P3	C3	0	Cus3	İSTANBUL	ESENYURT		Avrupa_2	3.120276214
P4	C4	0	Cus4	İSTANBUL	SANCAKTEPE		Anadolu_4	14.49156068
P5	C5	1	Cus5	İSTANBUL	GÜNGÖREN		Anadolu_2	6.973593088
P6	C6	1	Cus6	İSTANBUL	GÜNGÖREN		Anadolu_2	6.973593088
P7	C7	0	Cus7	İSTANBUL	SANCAKTEPE		Anadolu_4	14.49156068
P8	C8	0	Cus8	İSTANBUL	SARIYER		Anadolu_3	15.3558501
P9	C9	0	Cus9	İSTANBUL	SARIYER		Anadolu_3	15.3558501
P10	C10	0	Cus10	İSTANBUL	ŞİŞLİ		Anadolu_2	6.973593088

Figure 3.8 Assigned Risky Points and Calcuted Distances to the Each Customer Address

Exploratory analysis (described in section) and data preparation steps are applied to the data collection. Risky points and distances have been identified for each customer. The next step is to use the data to assess the impact of these variables on the renewal model in insurance developed by using the General Linear Model (GLM).

3.2 METHODS

3.2.1 Cross Industry Standart Process for Data Mining (CRISP - DM)

Cros-Industry Process For Data Mining (CRISP-DM) methodology provides a defined and standard approach to follow steps while developing a data-centric project. CRISP-DM is a well-proven methodology to understand the data and develop business requirements using data sets related to the project. It is practical and flexible to understand the data. It helps understand the business requirement structurally when using analytics methods. It is the reference flow that runs

through almost every data analytics study. The CRISP-DM modules and their relations with each other are shown in figure 3.9.

The CRISP-DM model is a set of events describing the steps of what to do and how to do it while understanding requirements and applying development activities. While applying the steps, following each other is not a must. They can be repeated or check the previous one every time, or they can go back to the previous one to complete the understanding of the studied data. It does not cover all the alternative routes between the steps in the model. The main idea of the CRISP-DM model is to understand the requirements and apply the development activities carefully and entirely without any misunderstanding.

Figure 3.9 CRISP-DM Proces

3.2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) defines the steps data workers can follow; data analysts, engineers, and data scientists. Data preparation is the most complex and time-consuming data domain activity. EDA helps to clarify the study and shorten the overall process. It is an approach to help to have information data set itself, and it is a philosophy to understand the way of processing data. EDA covers techniques to make the process easier, and these techniques are primarily graphical;

1. Getting the maximum insight into a data set
2. Uncover underlying structure
3. Define and Analyze important variables
4. Analyze and identify outliers and anomalies
5. Test underlying assumptions
6. Develop parsimonious models
7. Determine optimal factor settings.

The EDA is not a well-structured step-by-step technique. It is an attitude about how the data workers approach the data set to understand the hidden value in the data set. Figure 3.10 shows

where the EDA takes place in the data science process. The EDA aims to extract the real-world value from the data by applying the practices described in the process.

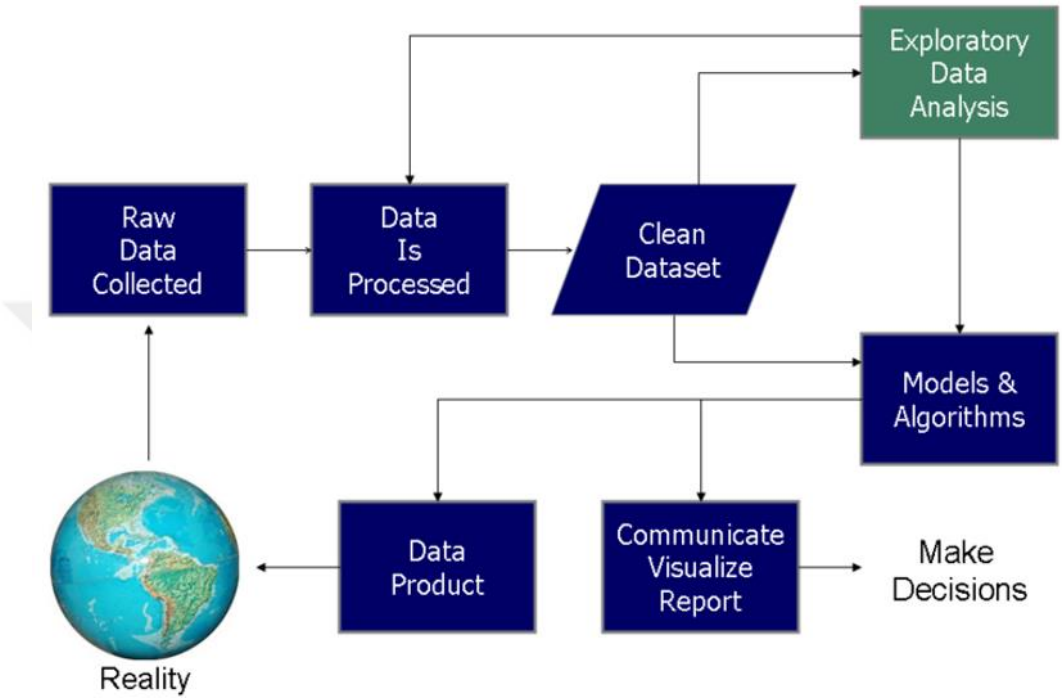

Figure 3.10 Data Science Process

EDA is not only statistical visuals or graphics, although these two concepts can be described almost the same. Statistical analysis and visuals are a collection of techniques that use graphics for all explanations and focus on only one data characteristic. EDA covers a more extensive scope while processing the data in the scope. EDA is an approach that puts all assumptions aside and focuses on analyzing the data with all the aspects that it can have. EDA is more than a technique but a philosophy of how the analyst anatomizes a data set about what we look for, look and interpret. Because of the steps or outputs of the process, there are similarities between “statistical graphics” and EDA, but EDA covers a vast scope and defines much more.

3.3 SOFTWARE

In this study, several technologies were used in the application and analysis process. In the following section, the details of the related technologies are mentioned.

3.3.1 Google Geolocation Application Programming Interface (API)

Google Maps API services interfaces enable programs to use google services. These API services are easy to connect, integrate, and get result for complex processes like gathering location information and helps developers to build simple applications for Web, iOS, and Android platforms. Users can get the altitude and longitude variables of a place inputted to the services using google API Key [51].

By integrating Maps API, developers can:

- *Show Google Maps* on the browser, iOS, or Android devices.
- *Place or pin a Marker on the map* when you want to indicate a specific geographic coordinate (latitude and longitude).
- *Show an Info Window* which is a popover so you can show more information about a place above the marker when clicked.
- *Draw a polygon* that covers a specific area on the map based on a number of coordinates specified in an ordered sequence.
- *Create a polyline* which is a path on the map based on a number of coordinates specified in an ordered sequence. The path line will be created between two lines, then the second to the third, and so on.

We used to place or pin a Marker API to get the latitude and longitude values from the customer address variables; city, county, district, and road.

3.3.2 Anaconda

Different languages help us to study data while developing data science projects. Python, R, and Julia are the most known programming languages that data scientists use. Python has been popular for working on data in almost every industry. R is mainly used in academia to perform statistical or machine learning algorithms on datasets. Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, etc.) that aims to simplify package management and deployment. Anaconda can run on all the operating systems with the help of the versions distributed, and they have different and prosperous data-science packages [48], which enable data scientists to develop, maintain and deploy their studies easily and quickly.

These four well-known components in anaconda distribution are PyPI, Conda, Anaconda Navigator and command-line interface (CLI). More than 250 packages come with anaconda installation, plus developers can access over 7.500 additional open-source packages from PyPI and conda packages. While working on anaconda distribution, developers run CLI and its GUI version Anaconda Navigator. All that a data scientist needs is defined, developed, and distributed with Anaconda development installation [49].

3.3.3 The Konstanz Information Miner (KNIME)

The Konstanz Information Miner is an analytical platform and modular environment that helps data analysts process data efficiently and enables visual assembly [50]. Due to its modular architecture, it is also easy to develop data pipelines that cover data preparation, data science, and data distribution capabilities. The KNIME platform was designed as a teaching environment at Konstanz university. It helps the students and researchers to collaborate and enables them to

integrate their algorithms. The KNIME platform enables researchers to share their studies by adding nodes that are reusable modules.

Data volume has exponentially increased in recent decades. Also, in academic research large volume of data is generated while simulating the studied topic. The researchers need easy-to-use and modified flows and then explore the results visually. The KNIME platform enables users to apply changes quickly and interactively. Plus, results can be seen and tracked easily. All requirements reviewed, the pipeline environment is the most suitable architecture. It allows the user who can develop modular and reusable components to visually assemble and modify the data analysis process from standardized components. In parallel, it offers an intuitive and visual way to document and share with the stakeholders what has been done.

Knime architecture allows valuable capabilities to the users and has three main principles:

1. *Visual, İnteractive Framework*: There are pre-built processing units. The flows are created to connect drag&drop units to process, modify, visualize, and distribute data.

2. *Modularity*: Thousands of predeveloped units with various usages can be used to develop data analytics flows. These units are independent components and can be used by connecting each other via defined gates. Each unit is encapsulated means there is no predefined type of unit. They can be used in any flow to process data for the user. The units can be modified, and the platform users can develop new units. Units can be defined as a unit collection to create a new reusable component.

3. *Easy Expandability:* Nodes the reusable components of the Knime platform, can be developed and added to the platform easily. It is as simple as the plug&play principle and does not need to install/deinstall steps.

In order to achieve this, a data analysis process consists of a pipeline of nodes connected by edges that transport either data or models. Each node processes the arriving data and/or model(s) and produces results on its outputs.

3.4 MACHINE LEARNING

Machine learning (ML) is a sub-branch, and data analytics focuses on artificial intelligence (AI), which focuses on processing mainly structured data using algorithms to mimic human learning. Training, testing and measuring accuracy are the main steps of the machine learning flows. Increasing accuracy to an acceptable level is the ultimate goal of developing data analytics using machine learning methods. Having increasing values on accuracy is the success factor, and machine learning developers try to have a good feature set to achieve acceptable accuracy.

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on using data and algorithms to imitate how humans learn gradually. There are many machine learning algorithms people use to solve problems and apply to the datasets they have [52].

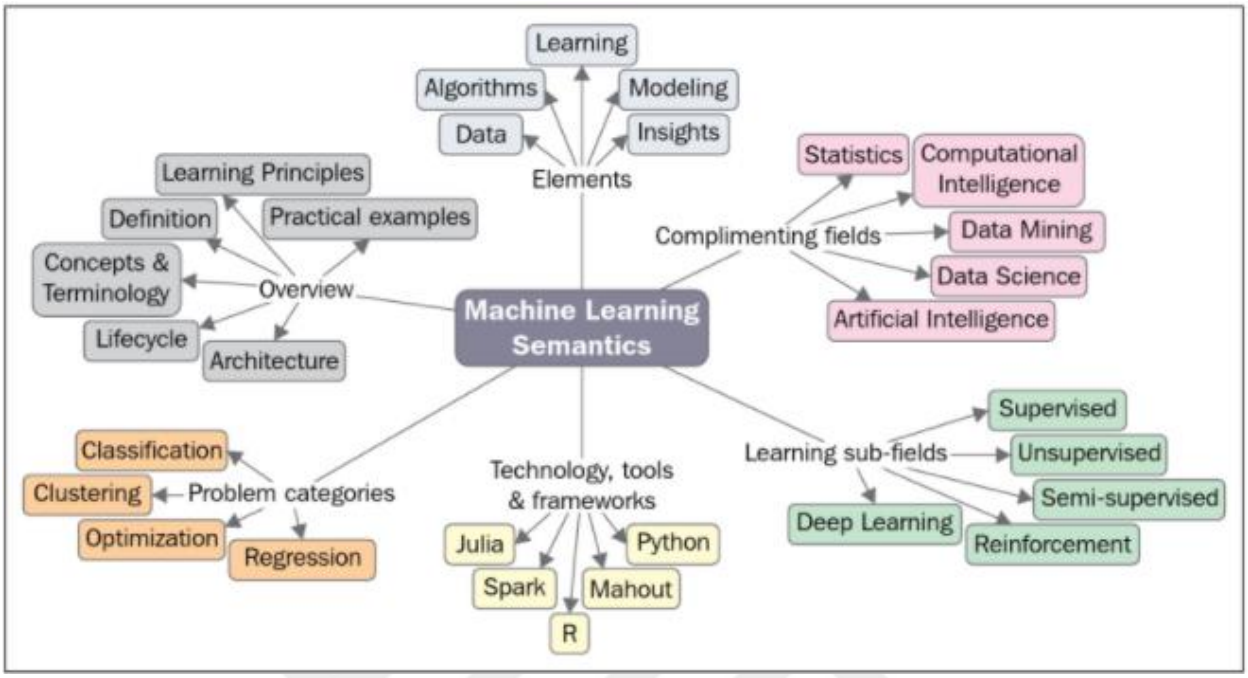

Figure 3.11 Key Concepts and Semantics of ML [53]

Data science is getting more critical while increasing data usage. Hence Machine learning is a crucial component of data science. Machine learning algorithms are used to make classifications to solve business problems or predictions to have insights into what will happen to the problem studied. The main idea behind applying machine learning algorithms to datasets is to get insights and solve the problem using data. These insights drive decision-making within applications and businesses, ideally impacting key growth metrics. ML learning flows are created using different methods that help users create solutions more straightforward and faster such as TensorFlow and PyTorch [54], [55]. There are three main parts in the learning systems of a machine learning algorithm;

1. **A Decision Process**: Machine learning algorithms are mainly applied to a prediction or classification problem. A set of data can be labeled or unlabeled; the algorithm identifies

the pattern between the variables and makes an estimation or classifies it to support the decision process.

2. **An Error Function**: An error function identifies the accuracy of the model. If there are known errors, the accuracy can be assessed, or based on the algorithm used, the study can identify the accuracy.

3. **A Model Optimization Process**: The accuracy o the model is defined as the success criteria of the developed analytics flow. During the training phase, the model learns the pattern in the variables. Then the flow uses the actual data set to have the solution or insight. At the first run, there can be a threshold on the accuracy, which can be minimized by repeating the process. The process defines the evaluation and optimization of the flow to reach an acceptable accuracy level.

Machine learning (ML) is an area of computer science focusing on machines' intelligence in performing human-like tasks. Machine learning is discussed under three subcategories: Supervised, unsupervised, and reinforcement learning.

There learning methods in machine learning concept;

1. *Supervised Machine learning* is a procedure that the usage of the processed dataset can define. The dataset is labeled to train the algorithm, and this labeled data set clarifies the training method of the flow and produces classified or predicted outcomes. Accuracy is the target success factor and is directly related to the labeled dataset used in the training step. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. The process needed to be trained and checked till the acceptable accuracy

reached the model to avoid overfitting or underfitting. Several methods are used in supervised learning, including neural networks, naïve Bayes, linear regression, logistic regression, random forest, and support vector machine (SVM) [56]

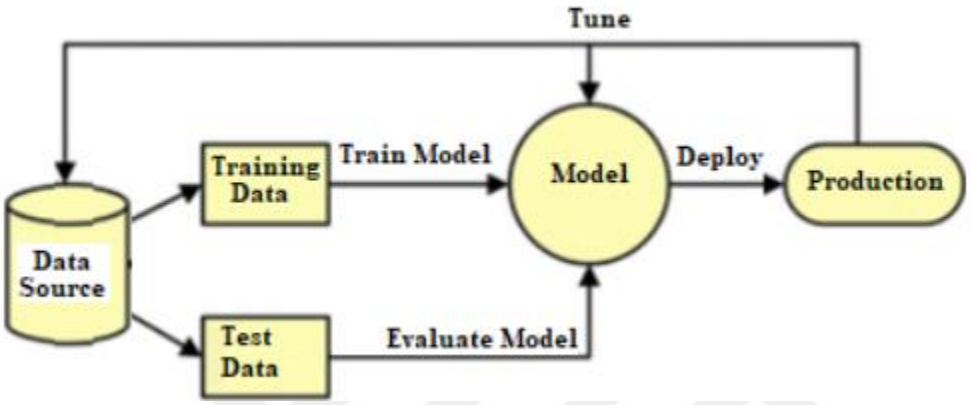

Figure 3.12 Supervise Learning Workflow [57]

2. *Unsupervised Machine learning* is distinguished from the supervised version from the dataset used. Unsupervised Machine Learning uses algorithms to analyze and cluster unlabeled datasets. The unsupervised algorithms were developed to discover hidden relations or motifs / patterns. The better capability for unsupervised learning is that there is no need to involve human intervention to identify hidden patterns and groupings in the dataset. This method is used to prepare the data to have better EDA and group it by not dealing with the parameters, image recognition and pattern identification. Dimensionality reduction is an excellent example of an unsupervised learning application area. Because the algorithms can discover hidden patterns, and if a variable is not relevant to the target, it can be excluded automatically. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches. There are various types of algorithms

used in unsupervised learning, including neural networks, k-means clustering, and probabilistic clustering methods.

3. *The Semi-supervised learning* method uses both capabilities which supervised and unsupervised learning have. It offers a happy medium between two methods. While training the data, supervised learning labels the small portion of the dataset that unsupervised learning algorithms show off to discover hidden patterns in the unlabeled part of the data. If the dataset is enormous and there is a small set of labeled data, and labeling is a costly development process, then semi-supervised learning is the path that can be followed [58].

Reinforcement learning is a learning model trained very similarly to supervised learning. The difference in reinforcement learning is in the way of training the algorithm; the reinforcement algorithm is not trained using a sample dataset. This type of algorithm learns by using trial and error, which means the algorithm has inputs and based on the output, there is a price for the correct outputs and a penalty for the wrong ones. The defined process reinforced the algorithm to learn [37].

3.5 GENERAL LINEAR MODEL

This section will describe the Linear Models and General Linear models. They are slightly different machine learning algorithms based on statistical methods. Especially General Linear models have different statistical components like single Linear Regression, Multiple Linear Regression, Anova, Ancova, Manova, Mancova, t-test and F-test. One more version is the

Generalized Linear Model (GLiM). GLiM allows residuals to use the exponential family of distributions instead of normal distribution.

3.5.1 Standard Linear Model

The standard linear regression aims to define the relationship between a variable that we want to predict the results (i.e., risk, renewal, fraud, claim) and independent variables, which are defined explanators.

The linear relationship between the target and independent variables can be formulated as;

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3 + \ldots +} + \beta_p x_{p +} \epsilon_i = 1, 2, \ldots., n \quad (3.1)$$

where β_0 stands for the intercept. $\mathcal{E}_i$ represents the error term, which can be interpreted as a random noise accounting for all underlying non-systematic effects contributing to the model's measurement error. In matrix notation, this relationship is given as:

$$Y = X\beta + \mathcal{E} \quad (3.4)$$

where $X\beta$ is called the linear predictor. Using $E(\mathcal{E}) = 0$, it follows that

$$E(Y \mid X = x) = X\beta. \quad (3.3)$$

The linearity constraint for linear regression only applies to the coefficients β, which means any non-linear relation among explanatory and response variables may still be modeled. Polynomial regression, which is a form of linear regression, fits a non-linear relationship between the value of x and the corresponding conditional mean of y.

3.5.2 Generalized Linear Regression

Consider that more than one linear regression model is generalized as a set of generalized Generalized linear models (GLMs). There are three different components of GLMs, which are not much different from linear regression models. They are slightly different from each other. GLMs are made up of;

- An output variable, Y, where all observations of this variable are assumed to be independently drawn from an exponential family distribution;
- A vector of k input variables, X1, X2, …, Xk; and
- A vector of k+1 parameters, b0, b1,…, bk, and a link function g(), which allow us to write g(E(Y)) as a linear combination of our input variables. That is:

where m = E(Y).

The purpose of our link function is to transform our output variable so that we can express it as a linear combination of our input variables which is not, to transform our output variable to normality, as is often mistakenly believed to be the case.

Depending on the probability distribution from which we assume our output distribution to be drawn, certain link functions are commonly used.

Distribution	Link Function
Normal	• Identify: g(m) = m
Gamma	• Negative Inverse: g(m) = -1/m • Log: g(m) = ln(m)
Poisson	• Log: g(m) = ln(m)
Binomial	• Logit: g(m) = ln(m/(1 – m))

Figure 3.13 GLM Distribution And Related Link Functions [60]

When specifying a GLM, it is, therefore, necessary to specify the output probability distribution function and the link function.

A standard linear regression model is a particular case of a GLM where we assume a normal probability distribution and an identity link.

As their name suggests, generalized linear models (GLMs) are a generalization of the standard linear regression models. These models are based on the normal distribution extended to the exponential class of distributions (e.g., the normal, Poisson, binomial and gamma distributions). The class of GLMs was introduced in 1972 as a general framework for handling a range of common statistical models for normal and non-normal data and can be seen as the industry standard for modeling the relationship between the response and predictor variables [60]. GLMs in the insurance setting is thoroughly reviewed and applied in many models, which are currently very popular [61], [62], [63]. Three components define a GLM: a random component that specifies the probability distribution within the exponential family for the response variable Y; a systematic component that relates the parameter η to the predictors X, which is called the linear predictor $\eta = \beta X$; and a link function g(.), which must be monotone and differentiable, that connects the random

and systematic components [64]. The relationship between the dependent variable and its predictors can be represented as follows

$$g(\mu) = n_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} = \beta_0 + \sum \beta_j x_{ij}\,; \quad I = 1,2,\ldots,n \tag{3.4}$$

where the intercept is denoted by β_0, the explanatory variables by xi1, ..., x_{iz} and its coefficients by β_{i1}, ..., β_{iz}. Hence, the mean of the response variable can be written as:

$$\mu_i = E(Y_i) = g^{-1}(\eta_i) \tag{3.5}$$

Therefore, the two principle distinctions between the generalized linear model and the ordinary linear regression are:

1. In the linear model, the mean is a linear function of the explanatory variables, while in GLMs, “some monotone transformation of the mean. A comparative analysis of statistical models for the pricing of health insurance. Link function g(.), is a linear function of the explanatory variables with linear and multiplicative models as special cases” [65].

2. In GLMs, the response variable can follow any distribution that belongs to the exponential class of distributions, while in the linear model, the distribution of the response variable is restricted to the normal distribution.

In the insurance domain, researchers prefer to use a generalized linear model instead of ordinary linear regression for several reasons;

1- Linear regression models assume that a normal distribution is almost not fulfilled, as the response variable (severity and frequency) tends to have other distributions.

2- Pricing models developed for insurance uses multiplicative models [66], which is more reasonable. It is said that, while studying claim severity data, ordinary linear regression models often perform well. On the other hand, in GLMs, standard linear models can be used as a particular case by defining an identity link $g(\mu i) = \mu i$ with a normal distribution.

3.5.3 The Components of the GLM

In a GLM, the outcome of the target variable is assumed to be driven by both a *systematic component* and *a random component*.

- The systematic component refers to that portion of the variation in the outcomes that are related to the values of the predictors. For example, we may believe that driver age influences the expected claim frequency for a personal auto policy. If driver age is included as a predictor in a frequency model, that effect is part of the systematic component.

- The random component is the portion of the outcome driven by causes other than the predictors in our model. The situation includes the "pure randomness"—that is, the part driven by circumstances unpredictable even in theory—as well as that which may be predictable with additional variables that are not in our model. As an example of this last point, consider the effect of driver age, which we describe above as part of the systematic component—if driver age is in the model. If driver age is not included in our model (either due to lack of data or for any other reason), then, from our perspective, its effect forms part of the random component. In a general sense, our goal in modeling with GLMs is to "explain" as much of the variability in the outcome as we can using our predictors.

4. RESULTS

This chapter explains the renewal model, which is developed using a general linear model and shares the results of the variables and how they are related to the target variable; customer policy renewal score.

4.1 LOCATION BASE RISK INSPECTION

The insurance industry values location variables very much. Primarily catastrophic claim management uses risk models which center the location information. Earthquake, flood, and landslide topics use geographical information systems to model the claim's impact or develop estimation models on environmental diseases. Location information is mainly used to transfer the view onto a map and combine customer, policy, and claim data with location information. As we examine in the study, accident reports have location variables, but this information is used to identify how the accident happened by the vehicle experts.

A researcher gets a dataset to calculate critical locations from an insurer in Tayland [59]. In the thesis, the researcher studied he got a set of accident data and extracted the locations. For each location, he created a risk value to identify how risky the location accidents happen after calculating the risk factor by adding sample route information and showing that the study is helpful for the drivers and insurers. In the data set, there is no policy information; the researcher stopped the study at this stage and offered some ideas as suggestions.

In this study, we moved further and created two new variables for the insurance industry, which can be very useful for actuarial models like risk estimation, claim management, renewal scoring, and price optimization.

4.2 POLICY RENEWAL MODEL

The insurance industry has various types of renewal models (Section 2.3). Insurance companies try to manage a customer base with fewer claim amounts and high renewal rates / scores. Renewal score has different values according to different policy types those insurers sell in the market. Health policies have higher renewal scores than other products, motor products are price sensitive, and renewal is vital for this policy type.

Some variables are used in the renewal GLM model. Table 4.1 shows a set of variables used to identify renewal scores. The primary purpose here is to get a higher renewal score by applying GLM Gama distribution and increase the number of customers who renew the policy with a profitable premium amount. The renewal model is also related to defining the correct price for the product served to the customer.

Table 4.1 GLM Renewal Model Variables

Renewal Model Factors	Description
Period Querter	The renewal period. Added as quarterly
Ccm	The motor capacity of the vehicle subject to the policy
Customer Type	Customers can be individual or corporate.
Dist Channel Type	The channel through how the insurer reaches the customer
Fraud Flag	Does the customer have any fraud case
Fuel Type	The fuel type of the vehicle
Campaign Policy	The policy is written in the scope of a campaign.
Auto Study Timeline	The overall timeline for the auto insurance policy for the customer.
Bank Specialist Flag	Insurance specialists are working in the banks for the insurers to write policies while the customer uses financial support.
Customer City-County	Customer city and county information
Number of Renewal	The number of the renewed policy.
Payment Deliquency	Customers payment habit
Payment Method	Customer payment method; cash, credit card etc.

Policy Type	The type of policy product that the insurer sells in the market
Policyholder Age	Customer age
Policyholder Gener	Customer gender
Policy Holder Marital Status	Customer marital status
Power of Vehicle	The power value ıf the vehicle engine
Renewal Premium	The premium amount of the renewed policy
Total Number of Offer	The total number of the offer shared with the customer
Total Paid Amount	The total money customer paid to the insurer for the policies
Vehicle Age	The age of the vehicle

4.2.1 Renewal Factor Analsysis (Location Related)

GLM tests the variable according to its formula (section 3.5.2). Insurers test a list of different variables to estimate the customer renewal score. The model tested location-related variables and estimated the factor value to understand how the final score was affected. These factor values are used in the actuarial model like risk estimation, price optimization, and claim estimation. Figure 4.1 shows us the renewal score distribution of the customers calculated according to the cities where they live. This view tells us; that there are different renewal scores for each city, and according to the customer portfolio that exists in that city, insurers define campaigns and marketing activities. For example, Mardin has the lowest renewal score; on the other hand, Tunceli and Eskisehir have the highest scores.

The insurer can focus on Mardin more than Eskisehir and activate more campaigns and marketing activities in Mardin. Plus, the insurer should be in touch with the customers living in Eskisehir. It does not mean that if a higher renewal score, the customer will

purchase the policy. There is always a possibility to churn, and insurers must define a relationship procedure according to renewal scores calculated via analytical models.

The city is the highest location variable determined in the model. Actuarial models consider the city variable as a risk factor. Based on the city in that customers live, risk factor changes. As it is seen renewal score also changes according to the city variable.

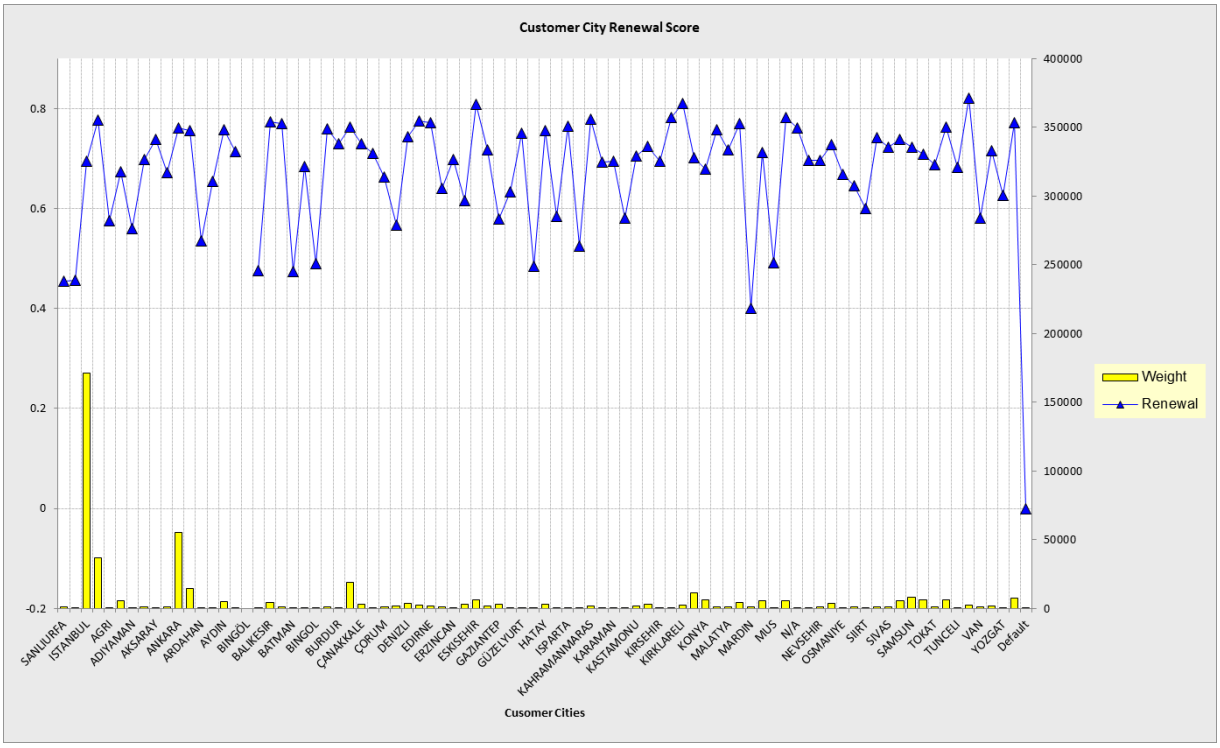

Figure 4.1 Customer City Variable And Renewal Scores

The county is the smallest unit of cities in Türkiye. Some counties have huge populations and relatively customers for insurers and claim amounts according to the number of customers living there. The county renewal values are not calculated for every city in Türkiye. Mainly, the scores are calculated as the same as the city of the county. However, some counties have different scores because of the area or population. Figure 4.2 shows the city-county pairs and their renewal scores. As is seen in Figure 4.2 the big cities like İstanbul, Antalya, Bursa, Ankara, and İzmir have county renewal scores. The county is another location-related variable. In Figure 4.1, Tokat has about 0.7 renewal scores, but in figure 4.2, there are less than more than 0.7 renewal scores for Tokat

counties.

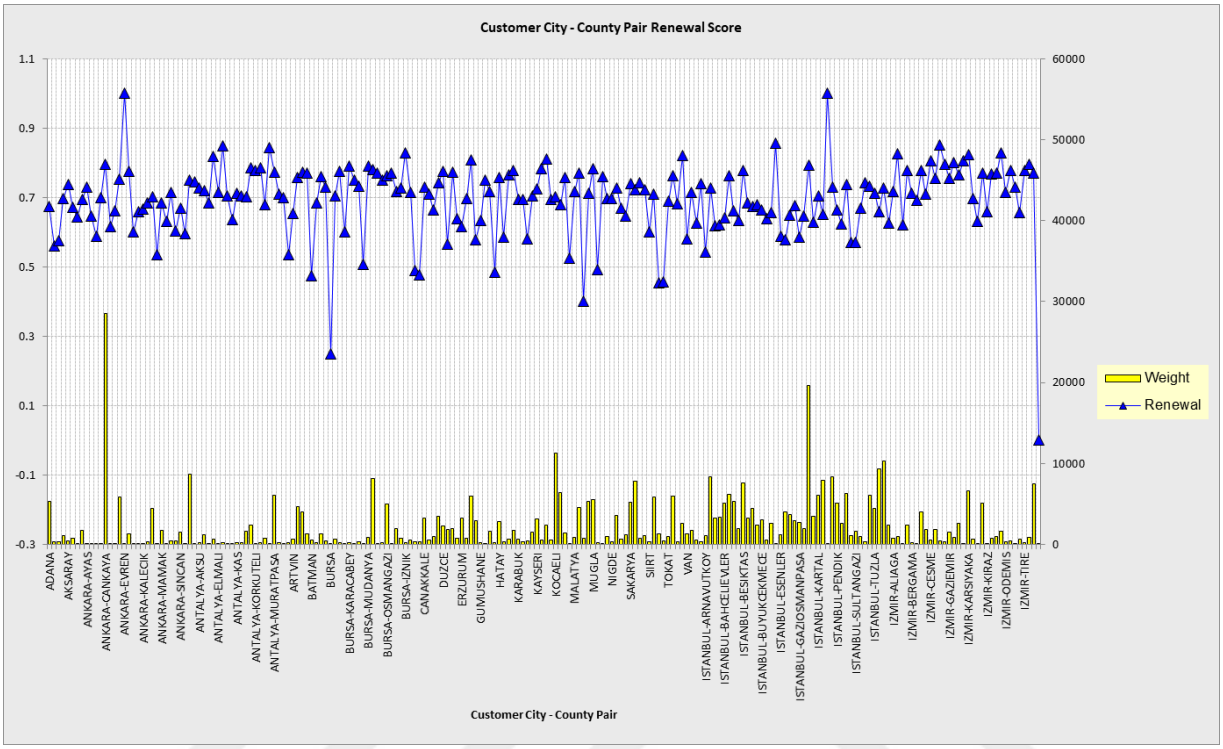

Figure 4.2 Customer City-County Variable And Renewal Scores

City-County pairs are essential for insurers because the number of vehicles is increasing every day. The city stays more general and aggregated to calculate the exact risk of the customers. Plus, claim management is challenging to manage at the city level.

Renewal score is also highly related to the number of customer claims. Claim means accidents for motor insurance. So, we can consider the number of accidents as a location-related variable. While calculating the risk level of the city and county, variable actuarial models check the total claim amount of the city and county. The number of accidents is more related to the customers. Figure 4.3 shows the relation between the number of claims variable and renewal score. Model output shows us that there is a decreasing slope between the variables tested with the GLM mode.

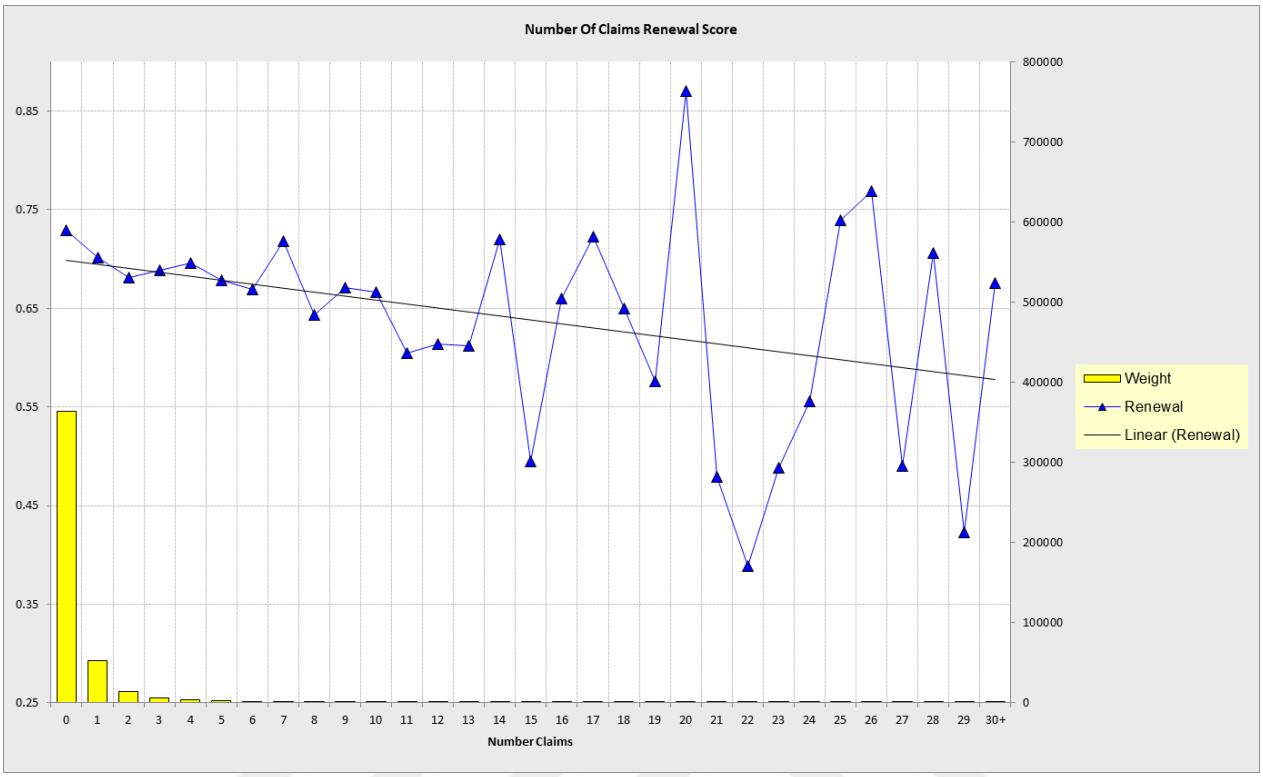

Figure 4.3 Number Of Claims Variable And Renewal Scores

In this study, we combine number of accidents that occurred at a point and city variables over accident reports by calculating the risky locations (see section 3.1).

Figure 4.4 shows the relation between the last claim location and renewal score. The last claim location can be necessary if there is a risk in the range of the customer's address. In figure 4.4, the mean slope has a decreasing angle on the renewal score. The last claim distance variable is in kilometer units. The variable tested in this view has a similarity with the thesis subject. We calculated the Euclidean distance between the nearest risky location and the customer's address.

As expected, the highest value on the renewal score is the no claim bar. If the customer does not claim, the insurer does not apply a significant increase in the policy premium. Motor insurance in Türkiye is price sensitive, and such price increases cause customer churn.

Actuary models try to find the optimum premium value for the insurer's customer base. The premium should be high enough to maximize the profit and low enough to save the customer. This is called price optimization, which is another popular actuarial GLM.

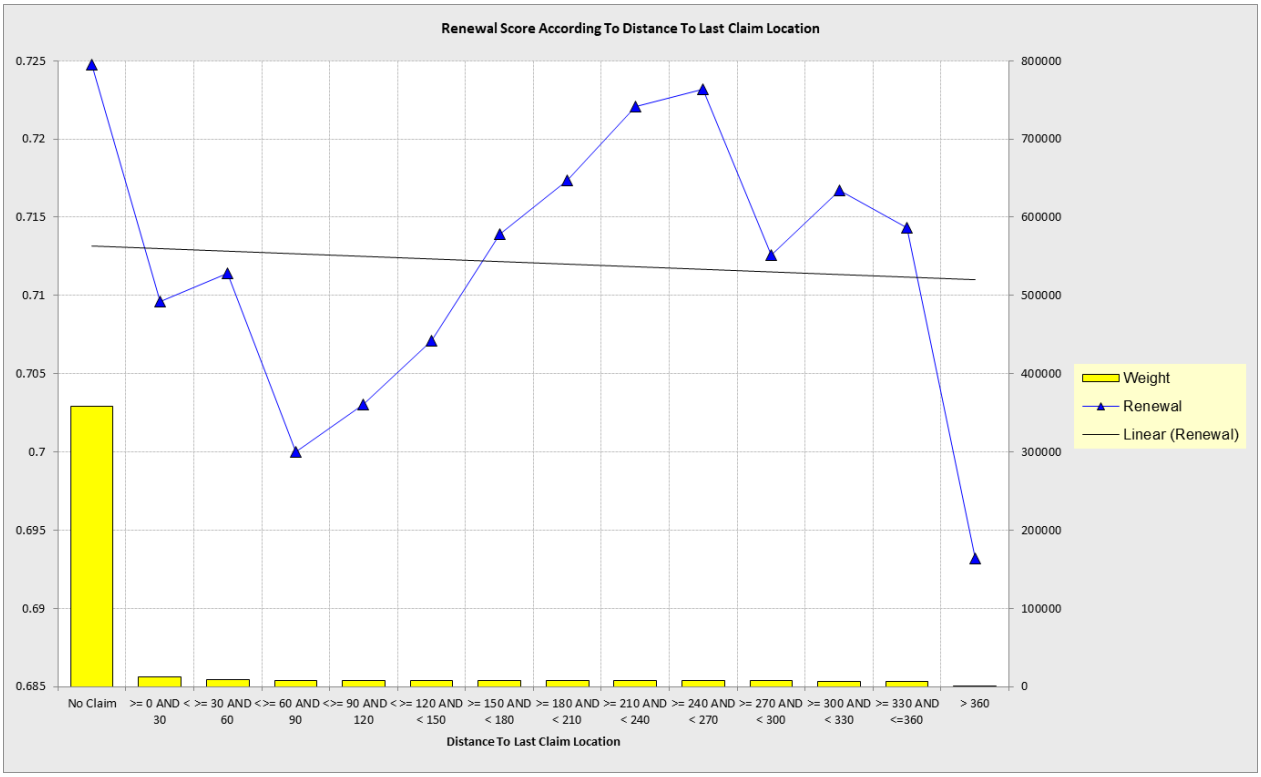

Figure 4.4 Distance to Last Claim Location and Renewal Score

4.2.2 New Variables Examanication on GLM

We studied two sets of data in this thesis;

1- The number of accidents at locations was collected from accident report data between 2011 and 2021. According to the amount of the number counted for the location, we defined four segments; small, medium, high, and urgent risky points. We visualize all the points on a map and get the latitude and longitude values of those locations. We decreased the data set by filtering İstanbul.

2- We prepared the policy list sold between Jan 2021 and June 2021. We checked their renewal value and again filtered the data set for İstanbul. Policy information supplies the address variables; city, county, district, and road. According to these address variables, we get each policy line's latitude and longitude values. Then the distance calculated between the customer location and urgent risky point location.

Knime analytics platform is the modular sequence-based low-code development environment defined in section 3.3.3. We processed the prepared data set with Knime and generated the renewal – risky point distance variable view shown in figure 4.5.

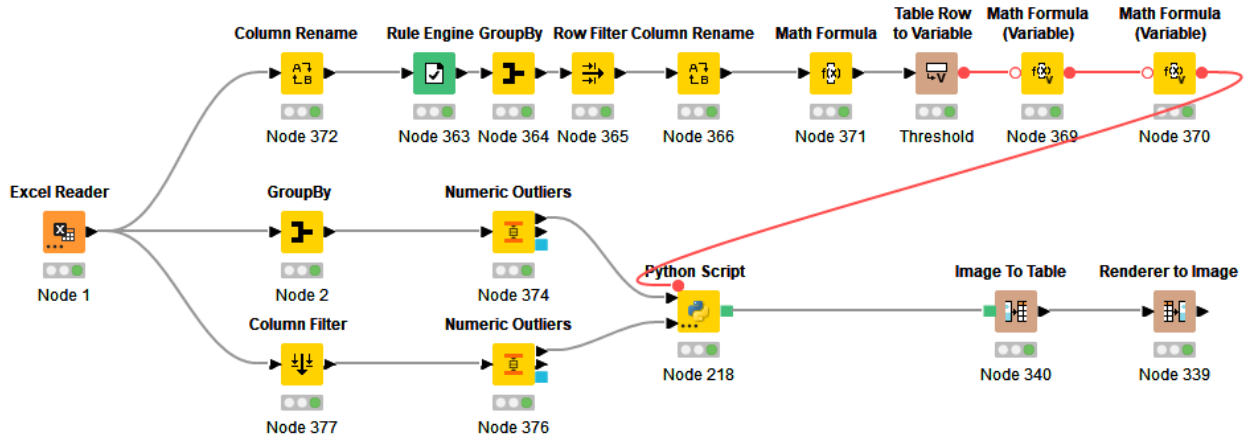

Figure 4.5 Knime Flow To Assess Distances To The Risky Locations on Renewal Score

The flow distributes the distances according to the number of policies assigned. The table of distance groups can be found in A.3.

The data set is processed into the flow. The outlier analysis is done, and the result can be seen in table 4.2. As is seen in the table, there is only one row that is not relevant to the process, and it is excluded from the data se.

Table 4.2 Outlier Analysis Result View

Outlier Column	Member Count	Outlier Count	Lower Bound	Upper Bound
Distance	844	1	-4.303	25.769

Figure 4.6 shows the GLM model output of the distance variable and renewal score. Each bar in figure 4.6 shows grouped distances and renewal score changes based on the distances.

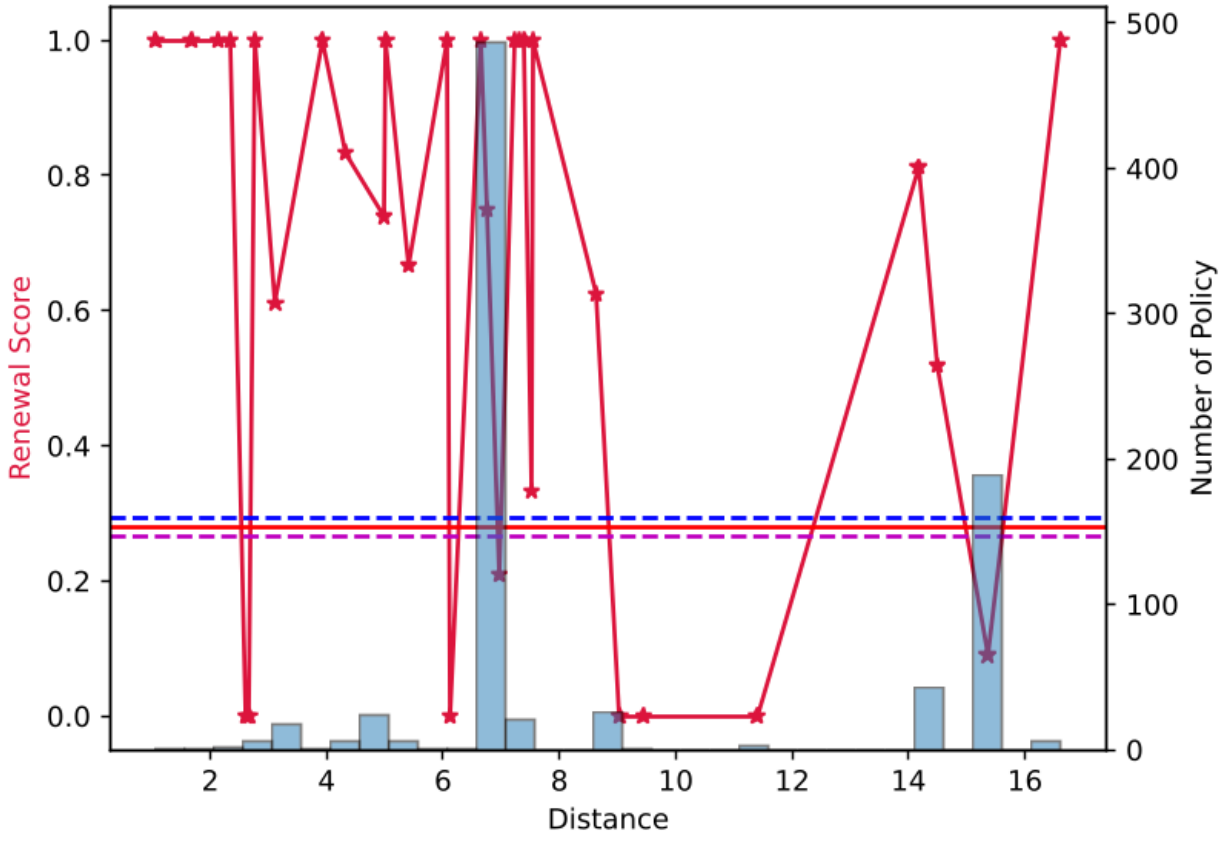

Figure 4.6 Risky Point Distance Variable and Renewal Score

Using the new variables in the GLM model for the renewal score estimation is highly possible, and as usual, insurance companies can create a factor multiplier for risky points and distances.

4.3 FINAL REMARKS

We processed the data and assessed the impact on the renewal score. The current renewal model processed four location-related variables; city, county, number of claims that occurred and distance to the last claim.

We created two new variables to enrich the GLM model for the renewal score estimations. As seen in Figure 4.6, the average renewal score changes based on the distance to the risky points. In conclusion, the calculated distance variable defines the risk, and if the customer lives away from the risky point, it has a different renewal score for the policy the customer purchased.

5. DISCUSSION

In this thesis, we analyzed accident reports in which the stakeholders prepare to define the accident and report to the insurance companies to request claim amounts within the scope defined in the insurance policy. We extracted all the accident report data from the history database and transformed it to clean and denormalize to prepare for the segmentation flow.

Risky locations were calculated based on the extracted data from the accident report database. Point by point, we counted the number of accidents that occurred. Based on the segments we defined in the study, we assigned four segments to each location accident occurred; small, medium, high, and urgent risky points.

The customer dataset also covers policy and claim information helped us to get location information of the customer addresses. The Euclidean distance calculation method helped us to get the distance between the customer address and the nearest risky point. Distance is the second variable that we created in this study.

In the renewal model flow developed with the GLM algorithm, many attributes have factor values that impact the price and risk models. Because, in the end, all the models related to continuous data are linear, insurance uses this very effectively. The newly created variables were added to the GLM flow and created a view showing how they are changing in the model in figure 4.6.

We can validate the result by controlling the claim amounts. The distance and claim risk is negatively correlated; when the distance increases claim amount decreases. Because if the

customer is far away from a risky point, there is a low probability of having an accident. Then customers do not have claim requests at the end.

These variables have factor values that affect other models used in insurance, and they can be a good candidate for those models.

5.1 FUTURE RESEARCH

During the study, we have encountered several opportunities that might be applicable, considering again in more detail in a future study.

We have connected the accident report location data and customer address information by calculating Euclidean distances. In a further study, the accurate distances can be found, and the risk accuracy can be increased. Insurance companies must invest in high-quality address data to study accurate distances, with complete data at least street level. A better way to have the distances is to find location information for each customer.

Another aspect of this study can be done on routes that pass through the risky points calculated by counting the accidents that occurred. If the customer can be tracked via telematics, the route information can be logged, and we can get how many times the customer passes over the risky points. The route information will also increase the risk point consideration and risk value accuracy.

In Türkiye, telematics is not a good way to collect data from insurance customers. Data privacy rules and data leakage examples make customers feel unsafe using telematics while driving. Instead, insurers can add value-added services to their mobile applications and push customers to

use them more frequently. Collecting a set of attributes via mobile applications installed on the customer's mobile devices is possible. Collected data also can be analyzed, and it is possible to create more variables for insurance services.

In this study, we examined the renewal model and defined the risky points and distance variables as inputs to the model. Insurance companies develop actuarial models. The pricing domain tries to estimate the best premium level for the customers, and there are main variable sets; customer and vehicle. By using the existing variables, insurers want to estimate the risk value of these two leading players. Only customer city, county and distance are considered the last claim variables, and location-based attributes can enrich by adding risky points and distance variables. When the insurer predicts the risk, optimizing customer-level premium calculation is possible.

Claim management is also vital for insurers to be profitable in the market. We locate the risky points on a map and share a visual view of the variables we created. Risky points can be used to predict the claim pool using geographical information system (GIS) methods. GIS has map base analysis, and by combining the location information with customer and vehicle data, claim management can gain more insight, and insurers have the potential to decrease the overall claim amount.

REFERENCES

[1] G. Dionne and others, *Handbook of insurance*, vol. 1119. Springer, 2000.

[2] N. Kuzieva, "Kuzieva Nargiza Ramazanovna Busness Processes In The Insurance System And Ther Features," *Архив научных исследований*, no. 24, 2020.

[3] P. Embrechts and M. v Wüthrich, "Recent challenges in actuarial science," *Annual Review of Statistics and Its Application*, vol. 9, no. 1, pp. 1–22, 2022.

[4] S. Haberman and A. E. Renshaw, "Generalized linear models and actuarial science," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 45, no. 4, pp. 407–436, 1996. .

[5] E. W. Frees, R. A. Derrig, and G. Meyers, "Predictive modeling in actuarial science," *Predictive modeling applications in actuarial science*, vol. 1, no. 1, 2014.

[6] P. J. Boland, *Statistical and probabilistic methods in actuarial science*. Chapman and Hall/CRC, 2007.

[7] S. Diacon, "A guide to insurance management," 2016.

[8] S. M. Coutts, "Motor insurance rating, an actuarial approach," *J Inst Actuar*, vol. 111, no. 1, pp. 87–148, 1984.

[9] W. S. Bato, "Improvement of non-motor claims processing of a non-life insurance company," 2017.

[10] S. Das and S. Kayal, "Ordering extremes of exponentiated location-scale models with dependent and heterogeneous random samples," *Metrika*, vol. 83, no. 8, pp. 869–893, 2020.

[11] D. Verzal, "What Does Data Science Mean for the Future of Actuarial Science?", 2018

[12] I. H. Sarker, "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, pp. 1–20, 2021

[13] S. Bressan, "The impact of reinsurance for insurance companies," *Risk Governance and Control: Financial Markets & Institutions*, vol. 8, no. 4, pp. 22–29, 2018.

[14] M. Brynjolfsson and L. Veldkamp, "A growth model of the data economy," *NBER working paper*, no. w28427, 2021

[15] E. Brynjolfsson, & B. (Eds) Kahin. "Understanding the digital economy: data, tools, and research", MIT press, 2002.

[16] R. L. Bornhuetter and R. E. Ferguson, "The actuary and IBNR," in *Proceedings of the casualty actuarial society*, 1972, vol. 59, no. 112, pp. 181–195Kuys, P. H. M. (1992).

[17] The actuary: From academic to professional. *Insurance: Mathematics and Economics*, *11*(2), 91-96.

[18] H. Paruchuri, "The Impact of Machine Learning on the Future of Insurance Industry," *American Journal of Trade and Policy*, vol. 7, no. 3, pp. 85–90, 2020.

[19] S. K. Singh and M. Chivukula, "A commentary on the application of Artificial Intelligence in the insurance industry," *Trends in Artificial Intelligence*, vol. 4, no. 1, pp. 75–79, 2020.

[20] A. E. Renshaw and R. J. Verrall, "A stochastic model underlying the chain-ladder technique," *British Actuarial Journal*, vol. 4, no. 4, pp. 903–923, 1998.

[21] M. Gesmann *et al.*, "Chain-Ladder." Paquete de R en Cran, versión 0.2, 2017.

[22] T. Mack, "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves," *ASTIN Bulletin: The Journal of the IAA*, vol. 21, no. 1, pp. 93–109, 1991.

[23] R. Olin, D. Vågerö, and A. Ahlbom, "Mortality experience of electrical engineers.," *Br J Ind Med*, vol. 42, no. 3, p. 211, 1985.

[24] J. Tomas and F. Planchet, "Constructing entity specific projected mortality table: adjustment to a reference," *European Actuarial Journal*, vol. 4, no. 2, pp. 247–279, 2014.

[25] J. Currie and H. Schwandt, "Mortality inequality: The good news from a county-level approach," *Journal of Economic Perspectives*, vol. 30, no. 2, pp. 29–52, 2016.

[26] I. D. Currie, "On fitting generalized linear and non-linear models of mortality," *Scandinavian Actuarial Journal*, vol. 2016, no. 4, pp. 356–383, 2016.

[27] R. Thatcher, V. Kannisto, and K. Andreev, "The survivor ratio method for estimating numbers at high ages," *Demographic Research*, vol. 6, pp. 1–18, 2002.

[28] R. Richman and R. Dorrington, "Estimating mortality rates when population and death data are both misreported–a statistical adaptation of the near extinct generations methods," *Age (Omaha)*, vol. 1, no. 02, p. 6, 2017.

[29] S. M. Coutts and R. Devitt, "The assessment of the financial strength of insurance companies by a generalized cash flow model," in *Financial models of insurance solvency*, Springer, 1989, pp. 1–36.

[30] G. Castellani *et al.*, "Machine learning techniques in nested stochastic simulations for life insurance," *Applied Stochastic Models in Business and Industry*, vol. 37, no. 2, pp. 159–181, 2021.

[31] B. Liu, "Uncertainty theory," in *Uncertainty theory*, Springer, 2010, pp. 1–79.

[32] J. Zhai, H. Zheng, M. Bai, and Y. Jiang, "An Uncertain Alternating Renewal Insurance Risk Model," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[33] K. Yao and X. Li, "Uncertain alternating renewal process and its application," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1154–1160, 2012.

[34] X. Zhang, Y. Ning, and G. Meng, "Delayed renewal process with uncertain interarrival times," *Fuzzy Optimization and Decision Making*, vol. 12, no. 1, pp. 79–87, 2013.

[35] L. Palán, M. Matyáš, M. Vál'ková, V. Kovačka, E. Pažourková, and P. Punčochář, "Accessing Insurance Flood Losses Using a Catastrophe Model and Climate Change Scenarios," *Climate*, vol. 10, no. 5, p. 67, 2022.

[36] C. Pugnetti and M. Seitz, "Data-driven services in insurance: Potential evolution and impact in the Swiss market," *Journal of Risk and Financial Management*, vol. 14, no. 5, p. 227, 2021.

[37] J.-P. Boucher and R. Turcotte, "A longitudinal analysis of the impact of distance driven on the probability of car accidents," *Risks*, vol. 8, no. 3, p. 91, 2020.

[38] S. Husnjak, D. Peraković, I. Forenbacher, and M. Mumdziev, "Telematics system in usage based motor insurance," *Procedia Engineering*, vol. 100, pp. 816–825, 2015.

[39] P. Handel *et al.*, "Insurance telematics: Opportunities and challenges with the smartphone solution," *IEEE Intelligent Transportation Systems Magazine*, vol. 6, no. 4, pp. 57–70, 2014.

[40] J. R. Jahnert and H. Schmeiser, "The relationship between net promoter score and insurers' profitability: An empirical analysis at the customer level," *The Geneva Papers on Risk and Insurance-Issues and Practice*, pp. 1–29, 2021.

[41] S. Buczkowska, N. Coulombel, and M. de Lapparent, "A comparison of euclidean distance, travel times, and network distances in location choice mixture models," *Netw Spat Econ*, vol. 19, no. 4, pp. 1215–1248, 2019.

[42] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, "Pairs trading: Performance of a relative-value arbitrage rule," *The Review of Financial Studies*, vol. 19, no. 3, pp. 797–827, 2006.

[43] B. Do and R. Faff, "Does simple pairs trading still work?," *Financial Analysts Journal*, vol. 66, no. 4, pp. 83–95, 2010

[44] M. Soleymanian, C. B. Weinberg, and T. Zhu, "Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers?," *Marketing Science*, vol. 38, no. 1, pp. 21–43, 2019.

[45] S. Davari, K. Kilic, and G. Ertek, “Fuzzy bi-objective preventive health care network design,” *Health Care Manag Sci*, vol. 18, no. 3, pp. 303–317, 2015.

[46] L. Vincent, “Exact Euclidean distance function by chain propagations.,” in *CVPR*, 1991, pp. 520–525.

[47] M. Mohibullah, M. Z. Hossain, and M. Hasan, “Comparison of euclidean distance function and manhattan distance function using k-mediods,” *International Journal of Computer Science and Information Security*, vol. 13, no. 10, p. 61, 2015.

[48] D. Rolon-Mérette, M. Ross, T. Rolon-Mérette, and K. Church, “Introduction to Anaconda and Python: Installation and setup,” *Python for research in psychology*, vol. 16, no. 5, pp. S5–S11, 2016.

[49] D. Rolon-Mérette, M. Ross, T. Rolon-Mérette, and K. Church, “Introduction to Anaconda and Python: Installation and setup,” *Python for research in psychology*, vol. 16, no. 5, pp. S5–S11, 2016.

[50] M. R. Berthold *et al.*, “KNIME-the Konstanz information miner: version 2.0 and beyond,” *AcM SIGKDD explorations Newsletter*, vol. 11, no. 1, pp. 26–31, 2009.

[51] W.-C. Hu, N. Kaabouch, H.-J. Yang, and X. Wang, “Location based services using html5 geolocation and google maps apis,” 2013.

[52] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, “What is machine learning? A primer for the epidemiologist,” *Am J Epidemiol*, vol. 188, no. 12, pp. 2222–2239, 2019.

[53] S. Gollapudi, *Practical machine learning*. Packt Publishing Ltd, 2016.

[54] M. Abadi *et al.*, “{TensorFlow}: a system for {Large-Scale} machine learning,” in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283

[55] N. Ketkar and J. Moolayil, “Introduction to pytorch,” in *Deep learning with python*, Springer, 2021, pp. 27–91.

[56] R. Sathya, A. Abraham, and others, “Comparison of supervised and unsupervised learning algorithms for pattern classification,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.

[57] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[58] I. Sutskever, R. Jozefowicz, K. Gregor, D. Rezende, T. Lillicrap, and O. Vinyals, “Towards principled unsupervised learning,” *arXiv preprint arXiv:1511.06440*, 2015.

[59] Y. G. Wu and B. Y. Hsu, “Insurance Assessment Based on the Big Data of Driver Historical Timeline and Car Accident Occurrence,” in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 2021, pp. 1–6.

[60] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models,” *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[61] R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit, *Modern actuarial risk theory: using R*, vol. 128. Springer Science & Business Media, 2008.

[62] P. de Jong, G. Z. Heller, and others, “Generalized linear models for insurance data,” *Cambridge Books*, 2008.

[63] M. Denuit and J. Trufin, “Effective statistical learning methods for actuaries,” 2019.

[64] R. J. Tibshirani, “Adaptive piecewise polynomial estimation via trend filtering,” *The Annals of Statistics*, vol. 42, no. 1, pp. 285–323, 2014.

[65] E. Ohlsson and B. Johansson, *Non-life insurance pricing with generalized linear models*, vol. 2. Springer, 2010.

[66] M. W. whoe and C. Murphy, “Event-related brain potentials to attended and ignored olfactory and trigeminal stimuli,” *International Journal of Psychophysiology*, vol. 37, no. 3, pp. 309–315, 2000.

APPENDIX A

A.1 DATASET

Latitude	Longitude
40.18852	29.06096
39.93336	32.85974
41.2797	36.33606
38.42373	27.14282
41.01823	29.12743
36.99141	35.33082
36.8121	34.64148
40.98333	29.12777
36.89689	30.71332
36.88275	30.76947
40.76544	29.9408
40.98187	29.05762
41.04564	28.82472
41.02123	28.77809
41.01894	29.05762
40.9498	29.17395
41.03431	28.66148
39.84518	32.92122
40.9975	28.85055
41.06861	29.02853
41.16229	29.04742
40.9184	29.22045
38.72048	35.48259
41.0814	28.98197
39.99042	32.69554
40.98319	28.8536
41.0287	29.29018
36.99347	30.689
37.06595	37.3781
41.06006	28.98811
40.89823	29.35987
40.88094	29.25774
41.055	28.93444
40.17943	29.02271
41.01686	28.94704
40.98479	29.34826
40.80251	29.43979
40.98019	28.72067
41.04815	28.90045

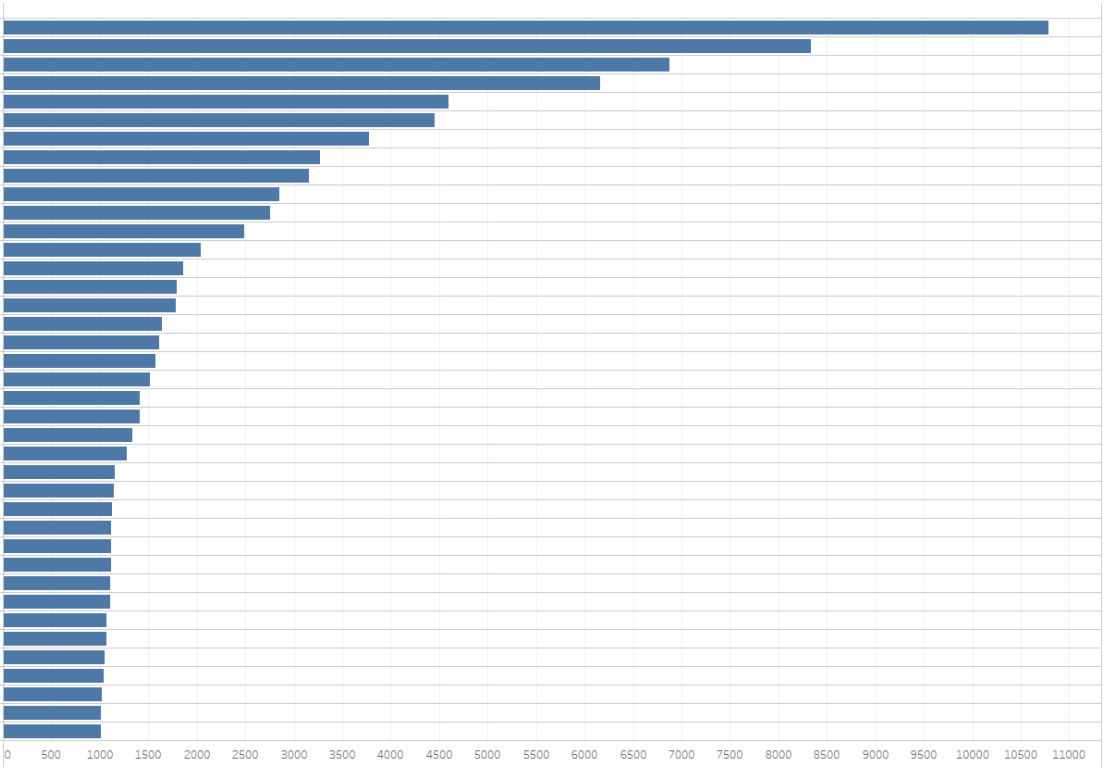

Figure A.1.1 Number Of Accident Distributions According To The Urgent Points

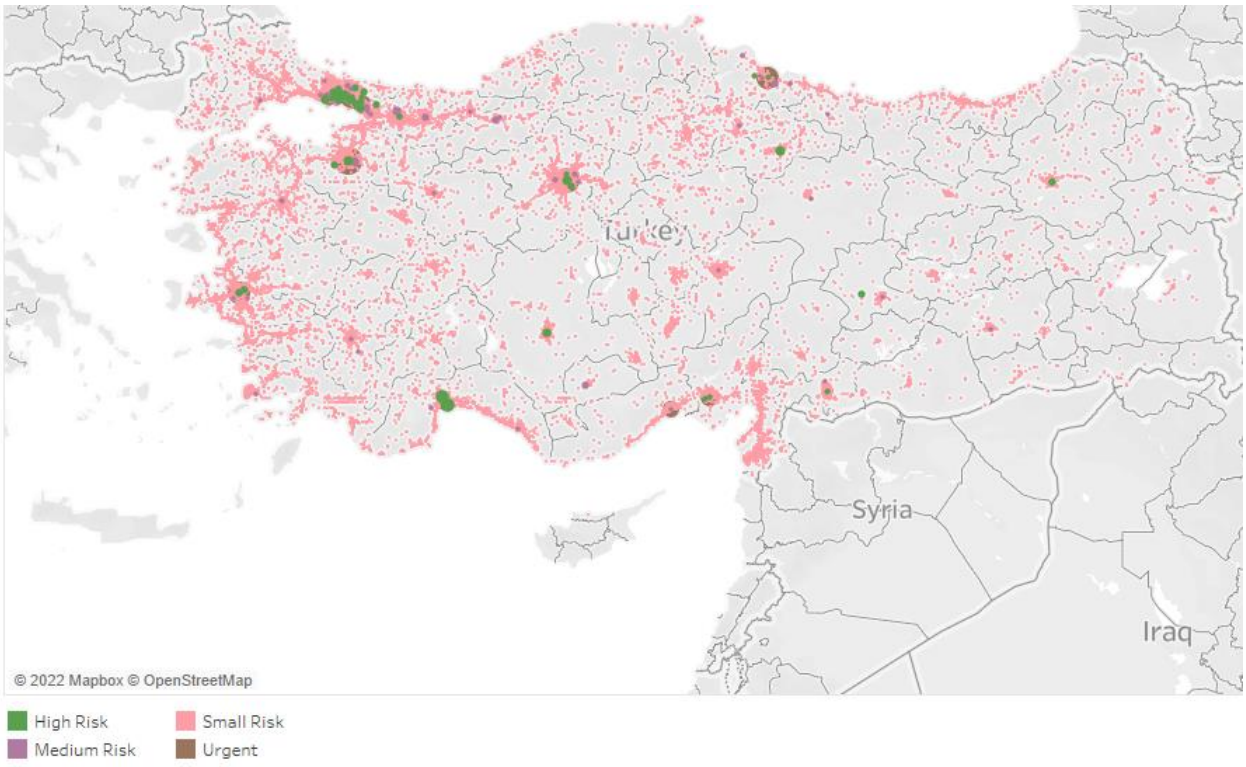

Figure A.1.2 All The Points Calculated In Türkiye Map

Figure A.1.3 All The Urgent Points Calculated In Türkiye Map

Table A1.1 All Urgent Points In Türkiye

Latitude	Longitude
41.03431	28.66148
40.98019	28.72067
41.02123	28.77809
41.04564	28.82472
40.9975	28.85055
40.98319	28.8536
41.04815	28.90045
41.055	28.93444
41.01686	28.94704
41.0814	28.98197
41.06006	28.98811
41.16229	29.04742
41.06861	29.02853
41.01894	29.05762
40.98187	29.05762
41.01823	29.12743
40.98333	29.12777
40.9498	29.17395
40.9184	29.22045
40.88094	29.25774
41.0287	29.29018
40.98479	29.34826
40.89823	29.35987
40.803	28.44
40.765	29.94
40.189	29.06
38.424	27.14
41.280	36.34
39.990	32.70
39.393	32.86
39.845	32.92
38.720	35.48
37.066	37.38
36.991	35.33
36.812	34.64
36.883	30.77
36.897	30.71
36.993	30.69

A.2 KNIME OUTPUT

Table A.2.1 The Policy Data Distributed According To The Auto Segmented Distance Intervals

Distance	Renewal Mean	Number of Policy At Same Distance
1.06221081	1	1
1.67527671	1	1
2.13838947	1	1
2.34760412	1	1
2.61894083	0	2
2.67196897	0	3
2.76448226	1	1
3.12027621	0.611111111	18
3.92152411	1	1
4.33043757	0.833333333	6
4.98936522	0.739130435	23
5.02340783	1	1
5.41103216	0.666666667	6
6.07333254	1	1
6.11960696	0	1
6.64262543	1	1
6.74990331	0.75	8
6.97359309	0.209205021	478
7.22477092	1	6
7.30627389	1	6
7.3993158	1	1
7.52238535	0.333333333	6
7.54605627	1	2
8.63710626	0.625	24
9.03812636	0	2
9.44113483	0	1
11.3949739	0	3
14.1660896	0.8125	16
14.4915607	0.518518519	27
15.3558501	0.08994709	189
16.6187803	1	6
464.4954	1	1