RETURN PREDICTION IN TURKISH STOCK MARKET VIA MACHINE LEARNING

A Thesis

by

Selen Babayakalı

Submitted to the

Graduate School of Sciences and Engineering

Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in the

Department of Financial Engineering

Özyeğin University

May 2022

RETURN PREDICTION IN TURKISH STOCK MARKET VIA MACHINE LEARNING

Approved by:

Asst. Prof. Emrah Ahi, Advisor,
International Finance Department
School of Business
Özyeğin University

Asst. Prof. Levent Güntay,
International Finance Department
School of Business
Özyeğin University

Asst. Prof. Erdinç Akyıldırım,
Department of Management
Boğaziçi University

Date Approved: 27 May 2022

To my dear family, colleagues and instructors,

with respects.

ABSTRACT

In this study I compare machine learning methods for predicting the stock returns of individual Turkish stocks listed in the Istanbul Stock Exchange (Borsa Istanbul). As the main machine learning model I use the Instrumented Principal Component Analysis (IPCA) and as a benchmark model I use Fama-French Factor Model. The IPCA model generates the stock-level expected returns based on observable stock-level and firm-level characteristics and latent common factors estimated within the model. Within the model stock-level characteristics determine the factor betas, namely the covariances of stock returns with the latent common factors.

I estimate versions of the benchmark Fama-French models between 3 to 5 factors. The versions of the IPCA models use between 3 and 6 factors and use 10 characteristics. The sample covers all stocks in the XUTUM Index and the sample period includes forecasts between 2010 and 2022. Using a panel data of 252 firms listed in the Borsa Istanbul XUTUM Index and I analyze the comparative performance of the IPCA and Fama-French models. More specifically, I look at the in sample and out of sample performances of the models by comparing the realized and predicted series of returns for each individual stock. I find that the IPCA model significantly outperforms the Fama-French model by obtaining significantly higher out of sample R-squared levels and correlation of return forecasts and realized returns. The performance difference between Fama-French and IPCA models is more pronounced in the Turkish stock market compared to results of [1] for the US stock market. Therefore, my results imply that the use of asset pricing models based on machine learning techniques may provide better results in emerging stock markets.

ÖZETÇE

Bu çalışmada, İstanbul Menkul Kıymetler Borsası'nda (Borsa İstanbul) işlem gören Türk hisse senetlerinin hisse senedi getirilerini tahmin etmek için makine öğrenimi yöntemleri karşılaştırılmıştır. Ana makine öğrenimi modeli olarak Enstrümanlı Temel Bileşen Analizini (IPCA) ve kıyaslama modeli olarak Fama-French Faktör Modelini kullanılmıştır. IPCA modeli, model içinde tahmin edilen gözlemlenebilir stok düzeyi ve firma düzeyi özellikleri ve gizli ortak faktörlere dayalı olarak stok düzeyinde beklenen getirileri üretir. Modelde stok düzeyi özellikleri faktör betalarını, yani hisse senedi getirilerinin gizli ortak faktörlerle kovaryanslarını belirler.

Hisse senedi tahmini karşılaştırmalı analizi için 3 ila 5 faktör Fama-French modelleri kullanılmıştır. IPCA modelinde 3 ila 6 faktör kullanılmış ve 10 karakteristik kullanır. Örneklem, BIST TUM Endeksindeki tüm hisse senetlerini kapsıyor ve örnekleme dönemi, 2010 ile 2022 arasındaki tahminleri içeriyor. Borsa İstanbul BIST TUM Endeksi'nde yer alan 252 firmanın panel verilerini kullanarak IPCA ve Fama-French modellerinin karşılaştırmalı performansları analiz edilmiştir. Daha spesifik olarak, her bir hisse senedinin gerçekleşen ve tahmin edilen seri getirilerini karşılaştırarak modellerin örnek içi ve örnek dışı performanslarına bakılmıştır. IPCA modelinin, örnek R-kare düzeylerinden ve getiri tahminleri ile gerçekleşen getirilerin korelasyonundan önemli ölçüde daha yüksek elde ederek Fama-French modelinden önemli ölçüde daha iyi performans gösterdiğini bulunmuştur. Fama-French ve IPCA modelleri arasındaki performans farkı, ABD borsası için [1] sonuçlarına kıyasla Türkiye borsasında daha belirgindir. Bu nedenle, sonuçlarım, makine öğrenimi tekniklerine dayalı varlık

fiyatlandırma modellerinin kullanımının gelişmekte olan ülke borsalarında daha iyi sonuçlar sağlayabileceğini ima ediyor.

ACKNOWLEDGEMENTS

I would like thank to my formal advisor Assistant Professor Emrah Ahi and Assistant Professor Levent Güntay for their assistance in all phases of this thesis study. Also my colleagues; Director of Center for Financial Engineering Faruk Özbaş, Efe Şen, Ali Erata and Kemal Peştreli thanks for helps and advices…

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Introducing the factors that explain the stock returns in both global and local markets is one of the most challenging issues in finance. In this context, wide ranging factor analysis, prediction method research and data processing studies have been presented to the literature in order to succeed in individual stock or portfolio return predictions. While as the milestone of the finance literature CAPM (Capital Asset Pricing Model) that discusses return and risk relation, many research papers have articulated it with the argue increasing the number of factors with different models the better prediction results is possible. The main importance of an asset pricing model is its power of prediction. Many studies find that CAPM only uses market beta to predict returns and thus is not an adequate model to predict asset return. After this observation studies extend the CAPM framework such as Fama-French 3, 4 and 5 factor models [2] , [3] and present that the factors and models enhance stock return predictions. Again the [4] discusses the return predictability for in-sample (IS) and out-of-sample (OOS) return forecasts with major variables. While the study introduces an exhaustive list of factors, the evidence still show that the predictability obtained from linear factor model may be limited.

As a major innovation in expected return prediction [1] propose the Instrumented Principal Component Analysis (IPCA) model. They argue that the model differs from the Principal Component Analysis (PCA). While similar to PCA IPCA generates the latent factors from a panel of stock return series, the IPCA and PCA differ in identifying the exposures of return to latent factors. In IPCA static and dynamically estimated stock-

characteristics guide the model in determining the covariance (betas) of stock returns and the latent factors. R^2

I estimate versions of the benchmark Fama-French models between 3 to 5 factors. The versions of the IPCA models use between 3 and 6 factors and use 10 characteristics. The sample covers all stocks in the XUTUM Index and the sample period includes forecasts between 2010 and 2022. Using a panel data of 252 firms listed in the Borsa Istanbul XUTUM Index and I analyze the comparative performance of the IPCA and Fama-French models. More specifically, I look at the in sample and out of sample performances of the models by comparing the realized and forecasted series returns of each individual stock. I find that the IPCA model significantly outperforms the Fama-French model by obtaining significantly higher out of sample R-squared levels and correlation of return forecasts and realized returns. The performance difference between Fama-French and IPCA models is more pronounced in the Turkish stock market compared to results of [1] for the US stock market. Therefore, my results imply that use of asset pricing models based on machine learning techniques may provide better results in emerging stock markets.

This thesis study includes the analysis of IPCA model on Turkish stock market and comparison with classical stock return prediction model proposed by Fama-French multi factor asset pricing model and seeks for the characteristics factors of firms for capital assets' impact on expected return

CHAPTER II

LITERATURE REVIEW

One of the most significant studies of finance literature is the CAPM that propose and enhanced by [5] and [6]. CAPM bases on the modern portfolio theory that provided by [7]. Many researchers argue that the CAPM model uses only market facto and market beta for asset return prediction and therefore is insufficient for explaining the variation in asset returns. The studies propose multi-factor asset pricing models by exploring many new additional factors. For instance, [8] use portfolios instead of individual assets. The most widely cited model is Fama-French 3 Factor model proposed by [2] where the factors are factor mimicking portfolio returns based on market return, firm market capitalization and the book-to-market ratio. Another model by [9] use principal components analyzing methods. [10] provide the new extended five-factor version of the FF3 model and include factor returns sorted by size, value, profit and momentum and exhibits that the five-factor model shows better results than the CAPM and FF3. On the other hand, [11] present a new contrarian model that uses asset characteristics of portfolios which includes market capitalization, book to market and previous years' return characteristics and proves them presents better performance at explaining cross sectional expected return of stocks.

Recent research by [12] present the stock return and asset characteristics' synchronized movement. With the widespread of using large dimensional and with large number of factors that are used in prediction these studies, higher prediction power properties are observed. [13] argue that larger datasets outgrowth using LASSO and PLS,

better return prediction results. While the by [12] provides that a couple of principle components for different 15 portfolios make possible to predict return with small value of alphas, IPCA model argues that it may exhibit well property with selected portfolios but the model that use PCA does not shows these properties in all portfolios. Concisely, IPCA beating this contradiction while pricing individual assets while using firm characteristics and its latent factor loadings, is better in categorizing the alpha and systematic risk disparity precisely.

After all these contributions the large dimensional factor model analysis is getting complicated. When the number of factors and time series properties, also factor loadings, increases it is harder to obtain precise results for individual stocks so the estimation method and factors that are used needs more sophisticated analysis. In literature most of analysis occurred with statistical factors in example [14] has studies on large dimensional data in cross section(N) and time series(T) properties which is contrary to classical approaches, and exhibits the results are homoscedastic with large N and calibrated T. With a similar approach IPCA is proposed by [15] that helps incorporate the characteristic and conditional information of capital assets to estimation model. [15] provided using the latent factors in an easier way with observable factors in the large dimension of dynamic data structures via using instruments in inner analysis for macroeconomic global data. It also provides working with large content of data in analysis. In this thesis study IPCA method applied the local stock market data (XUTUM Index members) and conditioned with both observed and latent factors.

CHAPTER III

DATA

3.1 Sample Construction

All the data used in this study are obtained from data providers Bloomberg, Thomson Reuters Eikon and Rasyonet/EquityRT Market Analysis and Research Platforms. To acquire optimum observation in terms of time series length and number of firms, and also for cross-checking the accuracy of data these three source are used. The sample firms are from XUTUM index, includes all members of Bourse İstanbul and cover a time range of 2008-2022 which gives 403 companies. In the data processing phase the companies which I discarded because of large missing data properties for given time range.

The initial part of the analysis is data extraction, cleaning and preparing for models because of missing dates and data points for the necessary sample frequency and sample period. First, I analyze stock with missing price observations and keep only the stocks that have at least 5 years of consecutive price data since 2016. Firms, that have more than half of their observations missing are excluded from the sample. To construct a large panel data with the maximum level of firms and numbers of firm characteristics I choose the sample start date as January 2010.

3.2. Data for Fama-French Model Factors and Characteristics

First, for the XUTUM index, I obtain the data for time range 2008-2022 which gives 403 companies as benchmark. The most difficult part of the analysis is the data extraction. For this process first it is controlled that the missing values only for price data of each stock for the last 5 years (between 2016-2022) and companies that have missing data with 0.50 threshold are rejected.

To benchmark the results of IPCA, the Fama-French 3 to 5 factor models [2], [3], [10] are used. Same missing elimination process applied for return and 1 month lagged return data. In order to sort data in terms of size, value, profitability, weakness and investment behavior market cap, price to book, operating profit, investment ratio, book to market variables are obtained for each specific company. For SMB factor (Small minus Big) the market capitalization data, for HML factor (High minus Low) book to market, for RMW factor (Robust minus Weak) operating profit, and for CMA factor (Conservative minus Aggressive) investment ratio data is used.

I split the stock returns to 10% to 90% quantiles and calculate the factor returns in line with the method of original of study of Fama and French. SMB, HML, RMW, CMA are calculated according to this methodology. In SMB and CMA factors, I use the difference of small quantiles and larger quantiles. In HML, RMW factors, I apply the opposite version. And I derive all factors' 24 month rolling mean values in order to use in rolling regressions. I obtain the data set from Bloomberg, Thomson Reuters Eikon and EquityRT to acquire optimum observation in terms of time series length and number of companies. As a robustness check, I compare the data set from different vendors and

make sure of the consistency of information and scale units whether the variables are reported as thousands or millions of TL.

3.3. Data for IPCA Model Factors and Characteristics

In the data processing of factors used in the IPCA analysis, I employ the factors proposed in [1], [13] , [4] and factors from widely known research in asset pricing. The factors that I use are market beta, market capitalization, momentum, price to book, volatility, volume, market cap to asset ratio, profit margin, total debt to asset, net income to free cash flow, stock return interest rate beta and FX beta. The details of the calculation are available in the methodology part. Additionally, I use the EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) to net interest expense ratio to capture characteristics related the profitability of firms. Bid-ask spreads give information about the price spreads and liquidity of individual stocks. But I eliminate the EBITDA to net interest expense ratio because of the missing data parts higher than 50% of necessary term. I decide the optimal number of companies of in the database to be 250 set the sample starting date as 2010-01-01.

As mentioned in Section 3 factors definitions are represented in Table 1.

Table 1. *Fama-French Model Factor Definitions*

Book-to-Market Value	Book value divided by market value
Investment Ratio	12 Months change in the amount of total assets of the company
Market Return	XUTUM Index Return
Market Cap.	Market capitalization
Operating Profit	Ratio of firm's operating income divided by total equity
Price-to-Book	Price of stock return divided by book value
CMA	Conservative minus aggressive, stock returns sorted by investment ratio and calculated by %10-%90 quantiles' difference
HML	High minus low, stock returns sorted by book-to-market ratio and calculated by %10-%90 quantiles' difference
RMW	Robust minus aggressive
SMB	Small minus big**,** stock returns sorted by market capitalization and calculated by %10-%90 quantiles' difference
Stock Retırn	Monthly stock price return of individual stock
Treasury Bill	3-month risk free rate bond of Turkish Government

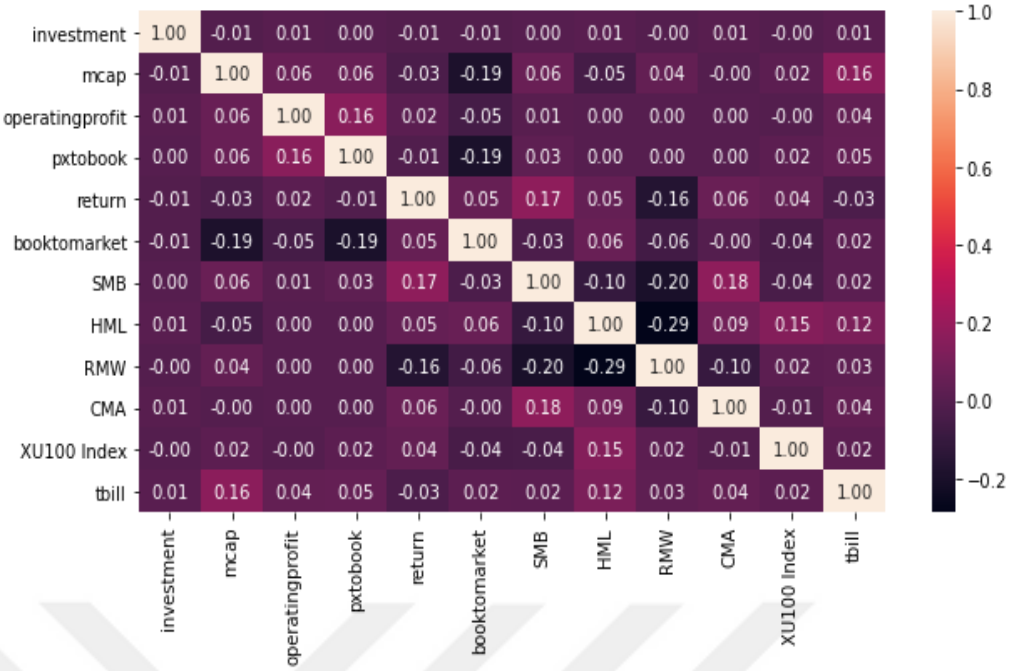

Figure 1.Correlation Matrix of Fama French Model Factors

In Figure 1 correlation matrix for Fama-French factors are given. The low even mostly negative correlation gives us chance to obtain better results in analyzes. As seen in figure the highest negative relation is between HML (higher book value firms and lower one's difference) and RMW which is profitability factor that may cause by the relation low valued firms profit ratio may have higher than high valued firms. Also for SMB and RMW relation may have same reason. But all given correlation results are suitable for analysis in model.

Table 2. *IPCA Model Factor Definitions*

Beta	The market beta of each individual stock
Book to Market Ratio	Book value divided by market value
FX Beta	USDTRY Exchange rate and 1 month lagged returns beta calculation for each stock
Market Cap.	Market capitalization
Momentum	Percentage change over last 6 months moving average price relative to benchmark index (source: Bloomberg)
Net Income to FCF	Ratio of the net income of firms in quarterly frequency to free-cash-flow value
Profit Margin	Firm's profitability ratio (Net income divided by revenue)
Return	Monthly stock return
Stock Return Int. Beta	Stock return interest rate beta, calculated with stock return and monthly interest rate beta
Volatility	Volatility in stock prices (source: Bloomberg)
Volume	Volume changes in stock prices, logarithm is applied for calculating change

Calculation methods for factors and definitions are represented in Table 2.

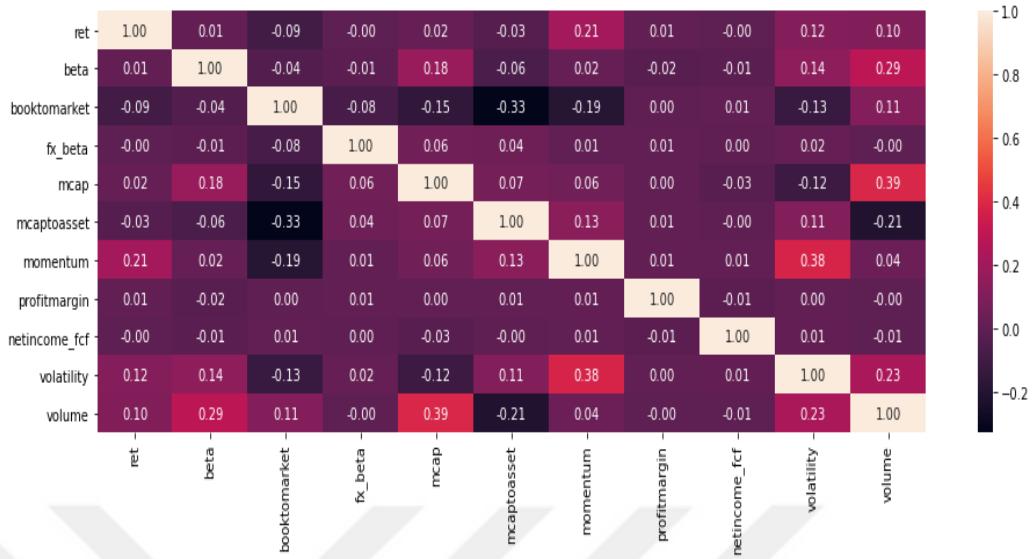

Figure 2.Correlation Matrix for IPCA Factors

In order to prevent multicollinearity and unbalanced results of model correlation matrices are pre-controlled before to use in model. The high correlation seen in dynamic factors which are volatility and momentum because of both are the reflection of the activity in the markets from these two factors. The second-high correlation seen between again the market beta and volatility it can be explained the market's dynamic movements' impact. The negative relation between market cap to asset value and book to market because both ratios consists market cap specifications. Briefly all correlations assumed acceptable and used in model. The diversification in factor correlations which consists positive and negative results are suitable for analyze and gives better results as you find in Empirical Results for Fama-French and IPCA Models section.

CHAPTER IV

METHODOLOGY

In this master thesis I used two models, one benchmark the other the actual model of this study.

4.1 Fama-French Model Methodology

On the part of Fama-French regression of study, I initially aim to analyze again the maximum time series range and maximum number of stocks. Both in IPCA and Fama-French model I use same equities and time in order to provide accurate comparison. I used in analysis beside the monthly stock returns also 1 month lagged stock returns. Market cap is used as a logarithmic form for calculation of size factor in order to provide smooth values. In order to calculate expected return, price to book in provided form for value factor, operating income and total equity which have quarterly frequency are grouped monthly and their division is used as operating profit for profitability factor calculation. To create panel data, I converted all factors via pivoting to use in IPCA model.

In order to get SMB (Small minus Big) factor, I ordered stock data by market capitalization and categorized to quantiles the bottom 10% quantile assumed as small stocks and top 10% quantile assumed as large stock and the average returns of each group extracted from each other. The same process applied and data sorted according to book to market values to get value factor HML (high minus low). In the part of profitability,

the operating profit margin, which is the percent change in 12 months of total assets, I used to obtain RMW (robust minus weak) and in the last part CMA (conservative minus aggressive) obtained by sorting data by investment ratio which is calculated by 12 months change in the amount of total assets of the company and used 3-month treasury bill in calculations as risk free factor. Also I used as market ß[beta) factor which is calculated with 24 months rolling window regression analysis. Variance Inflation Factor test measures the multicollinearity in a group of regression factor and I applied to all factors the VIF test to prevent multi-collinearity.

I used to analyze for Fama-French model XUTUM index and its 252 members which consist of Turkish Stock Market equities from 2010-2022. Initially the I aim the widest time series range and maximum number of companies for this study. Recent total number of XUTUM Index members is 403 as obtained from Bloomberg, but in the calculation of ratios which are not directly observed factors' missing values carry out misspecifications in model results. So after eliminate missing data points of current factors for each unique stock analysis continued with 252 stock and month-stock 36288 data points. Used monthly return, monthly market ß(beta) which is calculated with 24 months rolling window, market cap as size sorting factor.

24-month rolling regression is applied to factors by increasing the number of factors. Mkt-Rf (market excess) the difference between market return and 3-month bill of Turkey Government, SMB, HML, RMW and CMA used for a 12 months and 1-month monthly prediction. To get beta for each factor 24 months expanding and 24 months rolling window averages of these variables are analyzed. For both models the sectoral based, and size sorted R^2 results are obtained in analysis. Also, the profit and loss calculations are applied to data by sorting results by prediction error (difference between

actual and predicted) values and again categorized into quantiles then obtained as difference between large quartile minus small quantile which are used as 80% and 20%.

Table 3. *VIF Results for Fama-French Model Factors*

Book-to-market	1,10
Investment Ratio	1,00
Market Return	1,04
Market Capitalization	1,09
Operating Profit	1,04
Price to Book	1,06
Return	1,14
Lag Return(1 month)	1,09
Treasury Bill	1,14
CMA	1,05
HML	1,07
RMW	1,29
SMB	1,33

Table 3 represent the VIF results of Fama French factors. All values are acceptable lower than 2,5 thresholds. The threshold limits are obtained where [16] proposed.

4.2 IPCA Model Methodology

This thesis study involves the new method in literature proposed by [1] called Instrumented Principle Component Analysis(IPCA) that uses common factors and latent factors and while exposing the good sides, imposes biases. IPCA permits factor loadings to be moderately determined by characteristics that are observable and function as instrumental variables for latent factor conditional loadings. IPCA model constructs a

statistical link between expected returns and characteristics of assets. Moreover, it is coherent the classical asset pricing theorems which proposes risk premium is only driven by risk exposures. Through the encapsulating instruments in model, IPCA provides knowledge about returns and helps to enhance the prediction by precise factors. The main motivation for IPCA used in this thesis study, the risk exposures that commonly used in literature are using as an agent and observable statistical factors which is used in wide range of literature of asset pricing. The panel data which includes whole companies' time series and other factors exhibited as $X_{i,t}$, $ß_{i,t}$ and f_t are stands for factors and beta loadings and $\mu_{i,t}$ represents the errors in equation and also λ_t represents for risk price. Equation.1. below:

$$X_{i,t} = ß_{i,t} f_t + \mu_{i,t} \quad (1)$$

And the main excess return equation that IPCA represented as:

$$ß_{i,t} = c_{i,t}\Gamma + \eta_{i,t} \quad \text{and} \quad \lambda_t \quad (2)$$

$$E_t(r_{i,t+1}) = \frac{Cov_t(m_{t+1}, r_{i,t+1})}{Var_t(m_{t+1})}\left(-\frac{Var_t(m_{t+1})}{E_t(m_{t+1})}\right) \quad (3)$$

where $c_{i,t}$ stands for L (dimension of data) instrumental factors and $ß_{i,t}$ is again factor loadings, and the matrix of Γ is the LxK (number of factor) vector. The main constraint of IPCA is explained as the matching instruments to the factor loadings should be fix over cross sectional individual stocks and time series as mentioned in [15] . $c_{i,t}$ is the instrumented data and in equation. With the opportunity to accessibility and collection of big data the cognitive performance of return prediction increasing in time. At this knowledge IPCA provide the shaping the latent factors that consists is model in the line

of least resistance. In addition [17] argued that the relation between individual stocks and return relationship could be predicted with the contribution of dynamic factors inference from the analyzing the company's dynamic behavior of centralization.

Briefly all specifications of firm or stock, factors both includes statistical properties and observable dynamic factor loadings, also latent factors have an impact on return prediction. IPCA lends assistance to realize these operations in model. Despite the offering tolerance for using wide range cross section and time series of data, IPCA have rigid properties. The data which have fewer cross sectional or divergent properties will be insufficient for making precise prediction. On the other hand in [15] proves that IPCA which have restricted dimensions exhibits significantly (√N times) faster compared with PCA analysis that analyzes static loadings.

Panel data used for IPCA analyzes for 252 stocks between 2010-2022 has been addressed with a unique ID number before applying the model. For the outliers in data winsorization is applied with 0.05 percentage. R^2 results have been obtained with Spearman correlation, Pearson correlation and prediction errors for overall and individual stock base. Each time series factor is used in a model with 24 months expanding window and makes monthly predictions as Out of Sample results.

The IPCA model proposes formula below in Equation 3 for asset pricing (excess return):

$$r_{i,t+1} = \alpha_{i,t} + ß'_{i,t}\, f_{i,t+1} + \epsilon_{i,t+1} \qquad (4)$$

for cross sectional expected return prediction

$$\alpha_{i,t} = Z'_{i,t}\Gamma_{\alpha} + v_{\alpha,i,t} \quad (5)$$

$$ß_{i,t} = Z'_{i,t}\Gamma_{ß} + v_{ß,i,t} \quad (6)$$

the N number of assets in T periods and vector of latent factor K. Initially instruments in the process of prediction of latent factors via observable characteristics enable the build factor model for returns with extra data points. At this point IPCA exhibits different specifications from traditional models such as PCA, that predicts factors structure by only from return data of stocks.

In addition, the time varying parameters are instruments that provide the estimation latent factor loadings that tries to find conditional return facts. Γβ allows the analysis of a wide range of characteristics data points to minimize risk exposures. This process has an impact on dimension reduction and decrease in noisy risk factors via clustering them into the linear form. As mentioned above in the literature asset pricing models are built on these characteristic factors but in time series models the individual stock returns may have noisy results. As a strength of the model, IPCA uses stock betas as parametric functions making return prediction for each individual stock by the help of using characteristics.

You can find the Variance Inflation Factor (VIF) test results for the variables used in the analysis

Table 4. *VIF Results for Fama-French Model Factors*

Beta	1,08
Book-to-market	1,09
FX Beta	1,02
Market Cap.	1,39
Market Cap to Asset	1,22
Momentum	1,28
Net Income to FCF	1,01
Profit Margin	1,00
Return	1,08
Stock Ret. Int. Beta	1,04
Volatility	1,44
Volume	1,52

VIF results of IPCA Factors are represented in Table 4. All values are between 1 to 1,52 which are smaller than 2.5 threshold and suitable for using in model together. The threshold limits are obtained where [16] proposed.

4.3. Unrestricted Model of IPCA (Γα ≠ 0)

The unrestricted version of IPCA model provides for built a model that can explain the expected return with characteristics except the explanation that comes from systematic risk. The Unrestricted IPCA Model accepts the intercepts of model as linear function of characteristic factor loadings defined as in (Kelly, Pruitt, & Su, 2018). The formula mentioned in (7) provides the mean returns are not proxy of risk exposure where rt+1 gives the vector Nx1 that occurred by individual company stock returns Zt is a matrix

that gives the individual firms characteristics with size NxL and ε_{t+1} represents the residuals. Aim of this model lowering the sum of squared errors as composite.The unrestricted model uses with slight differences as $f\sim_{t+1} \equiv [1, f_{t+1}]$ and $\Gamma\sim \equiv [\Gamma_\alpha, \Gamma_ß]$. Unrestricted model provides the use of an increasing number of factors by including a constant.

Here in this study the unrestricted IPCA model is used which accepts $\Gamma \neq 0$. This type of IPCA model accepts the intercepts as linear aggregation of instruments where obtained weights described as Lx1 vector of Γα:

$$r_{i,t+1} = z'_{i,t}\Gamma_\alpha + z'_{i,t}\Gamma_\beta f_{t+1} + \epsilon^*_{i,t+1} \quad (7)$$

This model contains the cross sectional regression of "returns in excess of alpha" (Kelly, Pruitt, & Su, 2018) on dynamic factor loading like betas and represents the model how makes optimal panel diversity in expected return prediction. For many Γα and Γß that estimated in analyzing process. IPCA assumes $\Gamma'_{\boldsymbol{\alpha}}\Gamma_ß = 0_{1XK}$ where $\Gamma_ß$ and $\Gamma_{\boldsymbol{\alpha}}$ regressing and gives as result residuals that estimated by $\tilde{\Gamma}_{\boldsymbol{\alpha}}$

CHAPTER V

COMPARISON OF IPCA WITH FAMA FRENCH MODEL & RESULTS

5.1 Test Statistics for Comparing the Performance of Asset Pricing Models

In order to create benchmark for this thesis study the IPCA model compared with Fama-French model which includes only observable factors. In this research it is estimated 3 to 6 factor IPCA model for unrestricted version (model alpha is not assumed as 0). In order to compare results the R^2 statistics is used as:

$$Total\ R^2 = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - z'_{i,t}(\Gamma_{\propto}^{\sim} + \Gamma_{ß} f_{t+1}\right)}{\sum_{i,t} r_{i,t+1}^2} \qquad (8)$$

Indicates that the variance of return over dynamic characteristic factor loadings for time series and firm based specifications of results. R^2 used in compare also explains the how individual stocks and portfolio has impacted from systematic risk. As mentioned in literature part, the classical approach explains return by volatility of risk factor because of diversified structure by identifying it, but IPCA differs from these approaches by not categorizing dynamics of varying risk prices. Also, keeping risk exposure constant and using predictive task only in instrument section. For IPCA model that used in study have 10 both static and latent factors for 252 stock with 10 factor for 145 months. Parameters that used in observable variables as formulized as (Kelly, Pruitt, & Su, 2018) 252/

(10+145) times makes parameter estimation more than IPCA model, but IPCA realize this process in very efficient way. (nearly 95% lower number of parameters).

5.2 Fama French Model Results

Table 5.Result Table for Fama French Model

	Regression R^2	**3 Factor**	**4 Factor**	**5 Factor**
OOS	*Total Pooled*	0,003	-0,285	-0,360
	Time Series	0,004	-0,278	-0,354
	Firm Level	0,026	-0,278	-0,301
	Time Series Pearson	0,054	0,058	0,050
	Time Series Spearmann	0,038	0,032	0,023
	Firm Based Pearson	0,039	0,042	0,036
	Firm Based Spearmann	0,038	0,027	0,023

**Firm based regression: applied 24 months rolling regression for all unique stock/firm*
**Time Series based regression: applied 24 months rolling regression for all unique date from 2010-01-01 to 2022-02-28*
**OOS: Out-of-Sample*

Table 5. includes out-of-sample(OOS) results for 252 stocks in XUTUM index from 2010-01-01 to 2022-02-28. The 3 Factor model which is also proposed in

(Fama & French, 1993) consists Market-Rf (market excess return), size factor SMB (small minus big) and value factor HML (high minus low) gives better results in OOS 24 months rolling OLS regression. In 4 Factor model which both consists FF3 factors and as an addition the profitability factor RMW (robust minus weak) gives lower R^2 results compared with FF3. In the 5 Factors model both includes FF4 factors and additional CMA (conservative minus aggressive) which is obtained from the investment characteristics of firm also gives negative results and lower values compared with FF3 model. The reason for that may this 12-year time range did not enough to analyze firms' profit and investment behaviors is not sufficient to obtain these characteristics. Correlations both calculated for each unique data(month) and also for each individual stock based. In Table1 seen the time series based regression results exhibits better results compared to company based results. The market fluctuation and impact on portfolio could be observed from these results but in individual stocks the predictive power of Fama-French model seen insufficient.

5.3 IPCA Model Results

Table 6.IPCA Model Result Table

	Regression	**3 Factor**	**4 Factor**	**5 Factor**	**6 Factor**
	R^2	0,0378	0,050	0,0495	0,0495
OOS	*MSE*	0,0144	0,0142	0,0143	0,0143
	Pearson Corr.	0,143	0,151	0,156	0,159
	Spearmann Corr.	0,118	0,122	0,124	0,126

IS	R^2	0,0632	0,0622	0,0673	0,0667
	MSE	0,0108	0,0109	0,0104	0,0105
	Pearson Corr.	0,214	0,216	0,216	0,218
	Spearmann Corr.	0,172	0,176	0,177	0,177

R^2 results calculated as mentioned in IPCA Model
MSE: is mean squared error gives the difference between actual and predicted stock returns' squared.
Pearson Corr: Pearson Correlation, provides the linear relationship between actual and expected return results.
Spearmann Corr: Spearmann Correlation, difference from Pearson correlation is that it is non-parametric.

Table 6 contains in sample and out of sample result parameters of IPCA model. In order to obtain factors, I used for IPCA model also sources Bloomberg Professional Services, Thomson Reuters Eikon and EquityRT and the elimination of missing data settled in the sense of these factors both for IPCA and Fama-French model. Also to obtain these results from IPCA model, I concern while aim using large data-set for longest time series term but the optimal data provided for 12 years and 252 company. When we compared it with (Kelly, Pruitt, & Su, 2018) which use in model 12000 stocks and nearly 40 years range annual prediction, this model has quiet fewer data points but results present better results with this restricted dataset. But when I compare this results with Fama-French results that used same firm's data with observable factors, IPCA model performs better R^2 results, while 5 factor (same number with FF5) prediction gives negative results, IPCA have R^2

score 4,9% with 1% mean squared error (MSE) in Out-of-Sample results. Also I observe the stability and regular incremental behavior in IPCA model results that I use 3 to 6 factor for monthly prediction. In the in sample results the R^2 values that I observed 6 factor model gives slightly better results 6,7% compared with 5 factor model. It is observed again a slight downtrend in 4 factor model and this is consistent with (Kelly, Pruitt, & Su, 2018). In future research I aim to increase the number of factor that expands given characteristics may have positive impact on results, and apply this method on Emerging Market dataset.

5.4 Sectoral Results and Comparisons of Models

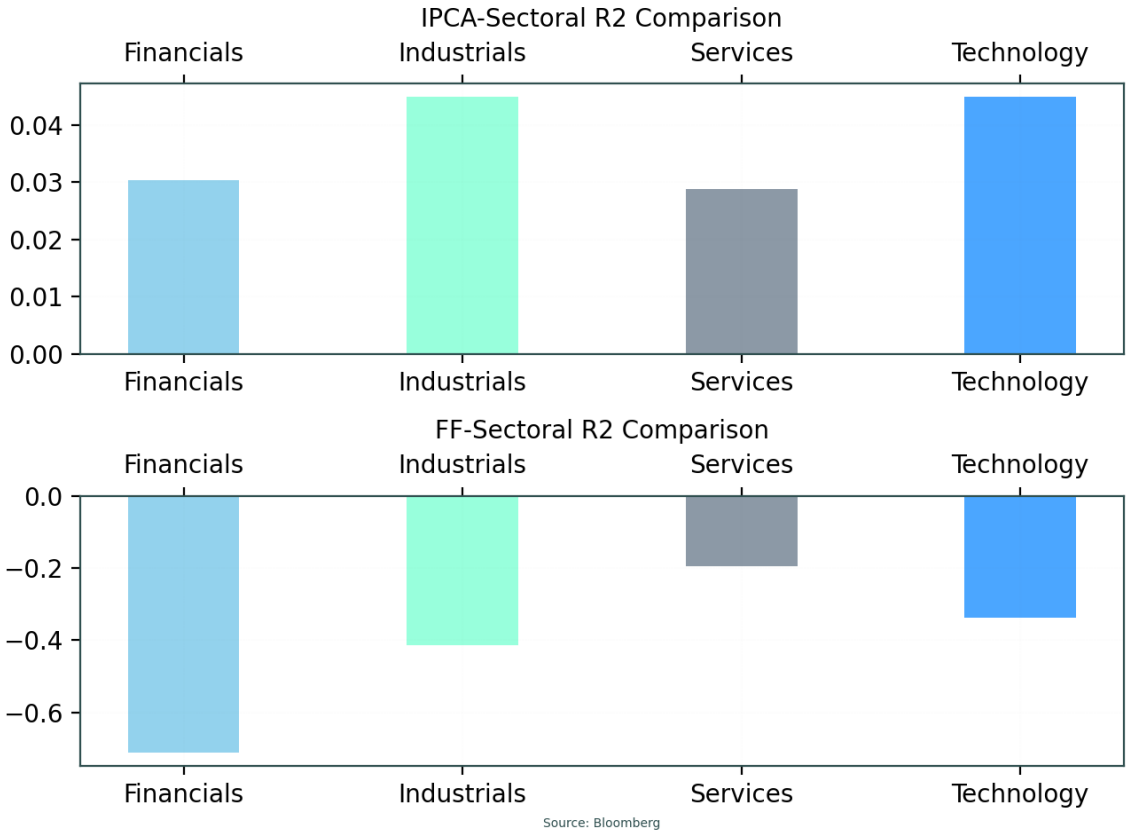

Figure 3 : Sectoral R Score Comparison of Models

Table 7.Sectoral Comparison

R^2 results	Financials	Industrials	Services	Technology
IPCA	0,03	0,04	0,03	0,04
Fama-French	-0,52	-0,41	-0,19	-0,33

In order to provide sectoral analysis, I receive the sectoral categorization information from Bloomberg for four main sectors which are industrials, services, financials and technology and they matched with IPCA and Fama-French model results. I compared the R^2 results according to sectors. In IPCA 6 factor model while technology and industrials sectors are leading the higher results with 4%, financials following it with 3% and services gives mean R^2 score of 2%. On the other hand, in the Fama-French model results are not valuable with negative results but services sector seen outperform others.

5.5 Small and Large Cap Analysis Results

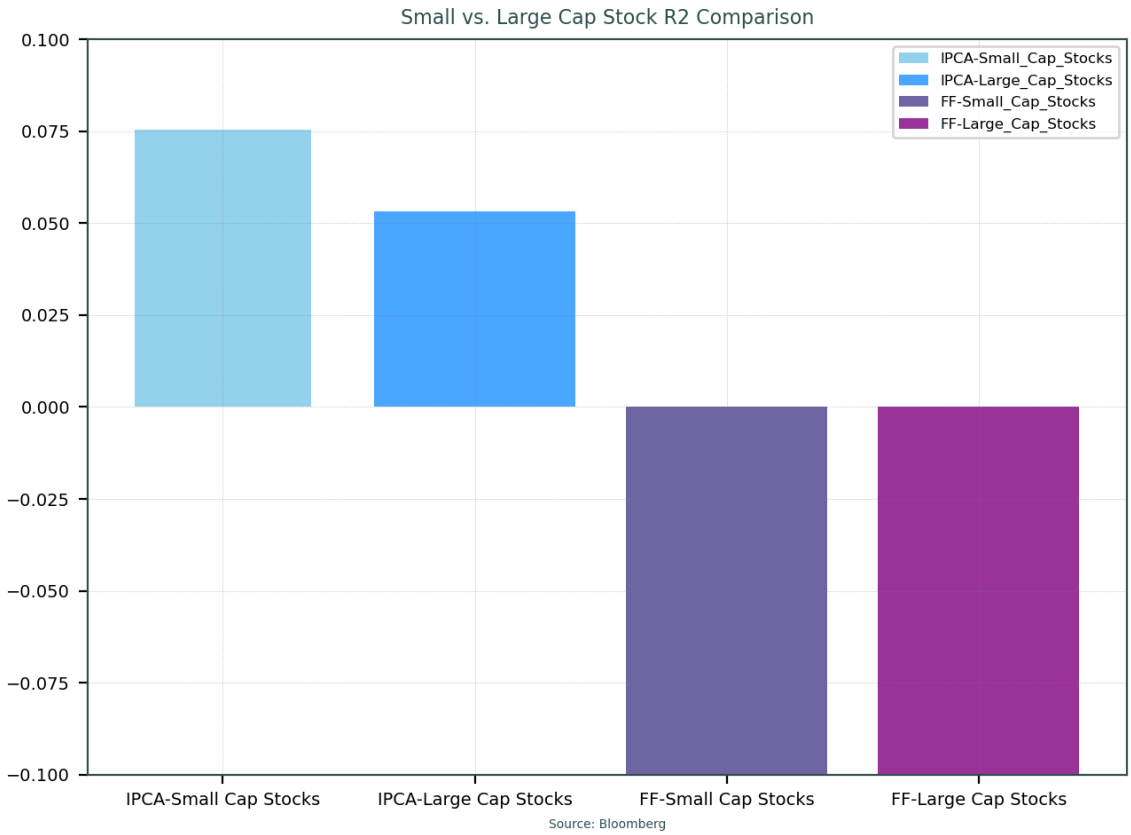

Figure 4. Small and Large Cap Firm Result Comparison

Used both for IPCA and Fama-French small and large cap analysis, sort the market capitalization values of firms and categorized into quantiles 10% to 90%. Small cap stocks placed in the top 10% quantile of IPCA data have higher R^2 results compared with large cap stocks. Both for IPCA and Fama-French model exhibits better results.

5.6 Time Series Based IPCA Analysis Results

As we can see in Figure 5. within increasing number of years also time series length of data increases and we can observe the R^2 score augmentation and decrease in prediction errors (calculated as difference between actual and predicted stock return

values). This information provides for future analysis with larger dataset may perform better prediction results.

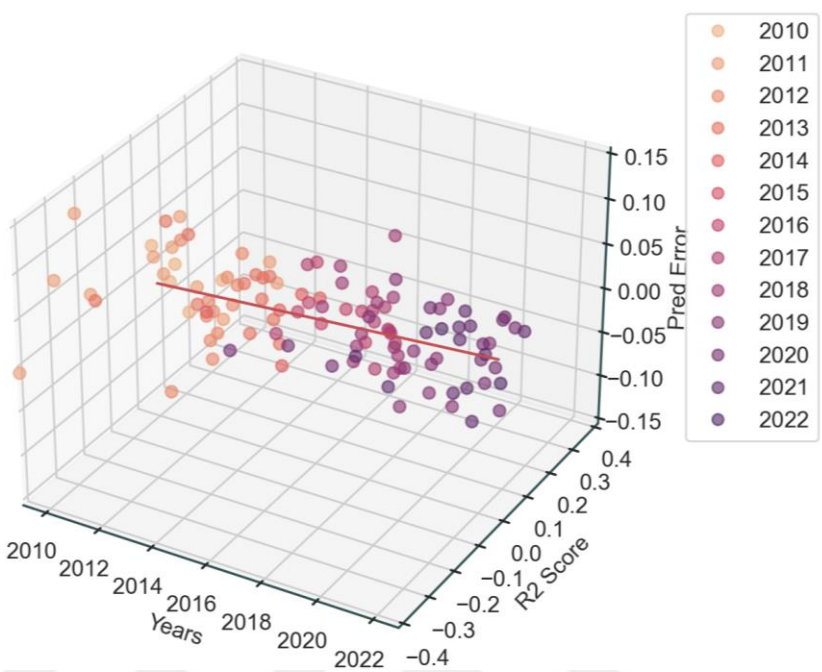

Figure 5. Time Series Based R2 Results

5.7 Firm Based Analysis Results

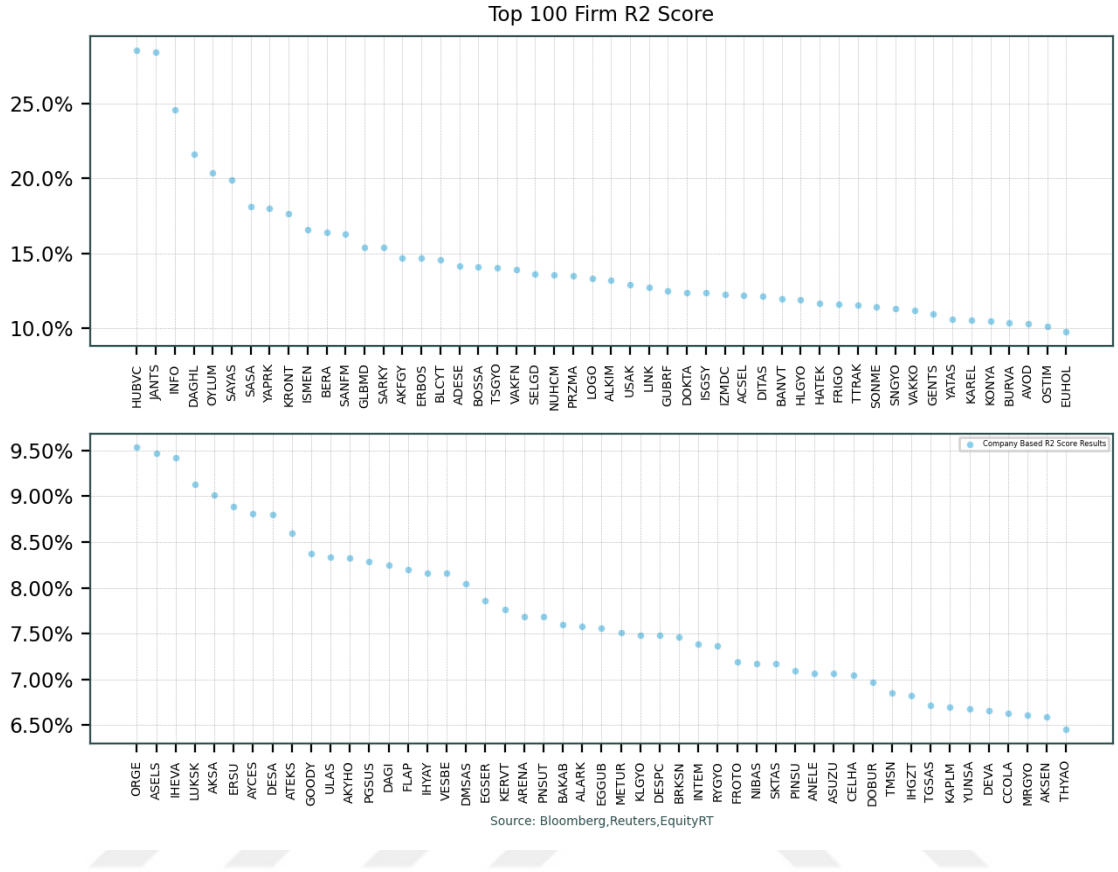

Figure 6. Firm Based R2 Score-Top 100 Firm

To exhibit the firm-based R^2 statistics variation for each specific stock, I split results into top and bottom 100 firms. In Figure 5 the results are shown and R2 changes between 28% and 6.45% for 100 firm with the highest R2.

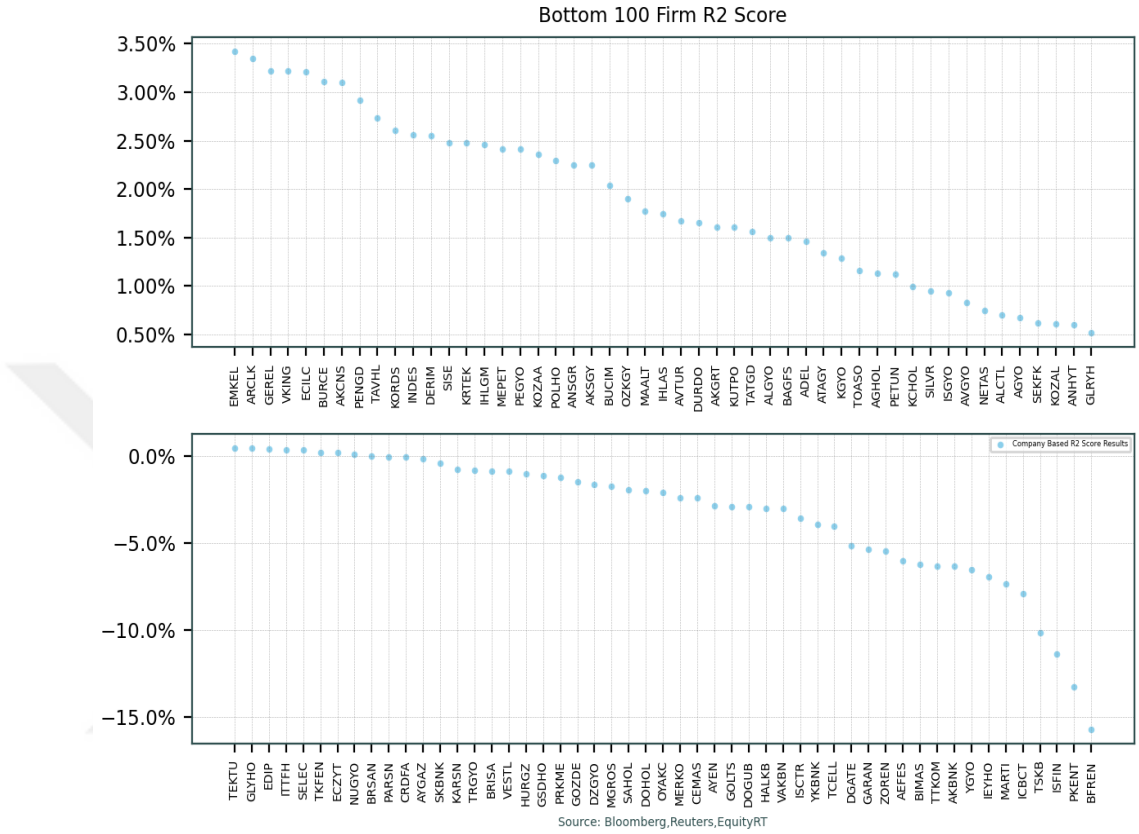

Figure 7. Firm Based R2 Score-Bottom 100 Firm

In Figure 7, you can see the results of bottom R2 ranked stocks, and R2 changes between -15.7% and 3.43%. The total average R^2 value for firm based results is 4% for the 6-factor IPCA model.

When I compare these results in Table 5 to IPCA model results in these figure IPCA model performs better compared to the firm-based model of the Fama French model.

CHAPTER VI

CONCLUSION

The main objective of this study is to predict stock returns for Turkish companies listed in the Borsa İstanbul via applying Instrumental Principle Component Analysis (IPCA). In the model I use observable stock-level and firm-level characteristics and latent common factors to obtain expected return forecasts. As a benchmark model, I use the Fama-French model. In first step, I obtain all factors for XUTUM Index members and process and transform them for the IPCA model. I require that at least 50% of observations are available for each individual stock and model factor during the sample period so that the stock or factor is included in the analysis. After missing data are discarded, we obtain a final dataset of 252 firms with 10 characteristics for the IPCA model. For the Fama-French model, I apply the same process for missing data and use same specific stocks, used 3 to 5 factors of Fama-French for a 24 months rolling regression as proposed in [3]. Also in IPCA's unrestricted method for nonzero alpha model, I apply the IPCA method for 3 to 6 number of factor analysis with a 24 month rolling window. Results of FF5's indicate that the model is insufficient to predict equity returns. Even FF3 gives positive and higher R2 values than FF5. In other words, measuring the profitability and investment factors which are used for FF5, larger dataset may exhibit better results. In IPCA model I observed that the increasing order in number of factors, the power of prediction also increases in 5 and 6 factor model. This research is achieved also when compared with [1]'s results. Briefly, IPCA model which uses both

static factors and dynamic factors with IPCA model beats the traditional approach of [3] which uses only observable factors.

To summarize, I find that the IPCA model significantly outperforms the Fama-French model by obtaining significantly higher out of sample R-squared levels and correlation of return forecasts and realized returns. The performance difference between Fama-French and IPCA models is more pronounced in the Turkish stock market compared to results of [1] for the US stock market. Therefore, my results imply that the use of asset pricing models based on machine learning techniques may provide better results in emerging stock markets.

REFERENCES

[1] B. Kelly, S. Pruitt and Y. Su, "Characteristics Are Covariances : A Unified Model of Risk and Return," *Journal of Financial Economics,* 2018.

[2] E. F. Fama and K. R. French, "Common Risk Factors in the Returns On Stocks and Bonds," *Journal of financial economics,* 1993.

[3] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of financial economics,* 2014.

[4] I. Welch and A. Goyal, "A Comprehensice Look at The Empirical Performance of Equity Premium Prediction," *Oxford University Press,* 2007.

[5] W.F.Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance1,* 1964.

[6] F. Black, "Capital Market Equilibrium with Restricted Borrowing," *The Journal of Business,* 1972.

[7] H. M. Markowitz, "Portfolio Selection Efficient Diversification on Investments," 1959.

[8] M. C. Jensen, F. Black and M. S. Scholes, "The Capital Asset Pricing Model : Some Empirical Tests," *Praeger New York,* 1972.

[9] G. Chamberlain and M. Rothschild, "Arbitrage, Factor Structure and Mean Variance Analysis on Large Asset Markets," *Econometrica,* 1983.

[10] E. F. Fama and K. R. French, "Size and Book to Market Factors in Earnings and Returns," *The journal of finance,* 1995.

[11] K. Daniel, M. Grinblatt, S. Titman and R. Wermers, "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *The Journal of Finance,* 1997.

[12] S. Kozak, S. Nagel and S. Santosh, "Shrinking the Cross Section," *Journal of Financial Economics,* 2017.

[13] J. Freyberger, A. Neuhierl and M. Weber, "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies,* 2017.

[14] J. Bai, "Inferential Theory for Factor Models of Large Dimensions," *Econometrica,* 2003.

[15] B. Kelly, S. Pruitt and Y. Su, "Instrumented Principal Component Analysis," *SSRN 2983919,* 2019.

[16] R. Johnston, K. Jones and D. Mabley, "Confounding and collinearity in regression analysis - A Cautionary tale and an alternative procedure," *Quality & quantity,* 2018.

[17] D. Acemoglu and P. D. Azar, "Endogenous Production Networks," *Econometrica,* pp. 33-82, 2020.

[18] S. A. Ross, "The Arbitrage Theory of Capital Asset Pricing," 1976.

[19] J. Lintner, "Security Prices, Risk and Maximal Gains from Diversification," *The journal of finance,* 1965.

VITA

After graduating from Mechanical Engineering with full scholarship at Özyeğin University in 2020, I started as a financial research associate at Özyeğin University, Center for Financial Engineering. I participate in research projects on Machine Learning with Asset Pricing, Big Data Analysis of Financial Data, Portfolio Optimization Algorithms, Measurement of Investor Risk Attitudes. In these projects, I collaborated with Turkish banks, regulatory organizations, brokerage firms and international agencies such as OECD. I am an M.S. candidate at the Financial Engineering and Risk Management Graduate Program, Özyeğin University.