Analyzing Individual Risk Taking Behavior with Machine Learning: An Application with a Pension Fund Data

Tolga. U Kuzubaş[1] and Burak Saltoğlu[2]

[1,2]Department of Economics, Boğaziçi University

November 28, 2022

Abstract

In this paper the actual risk taking behaviour of a private pension fund customers is studied based on 81,163 individual investors' account information. In our unique data set taken between 2018 and 2022 we have investigated the performance of risk profiling process. By using machine learning methods we analyzed how well the benchmark scoring methodology developed by the pension fund regulator fares in predicting actual risks. Applying the feature selection methods used in the machine learning literature we tried to select the most important factors behind the real world risk taking behavior . We found that that self-reported risk question, among other eight risk survey questions, seemed to be one of the most influential factors behind the individuals risk taken decision. Other survey questions, except for a simple lottery question, did not appear to be informative in determining the actual risks taken. Among the demographic factors, age of an investor turned out to be another crucial feature in our analysis. Moreover, we have observed that age also has a nonlinear impact on risk taking behavior. We have noticed that up to a threshold, aging is positive on risk taking behavior and then age begins to opposite direction with risk taking. Finally, we have conducted as prediction performance comparison between machine learning methods against the benchmark scoring model used in the industry . Many of the machine learning techniques, particularly boosting methods, had better predictive power than that of the benchmark. We conclude, machine learning techniques can improve the risk profiling process as a whole both as choosing the relevant variables and weigh them optimally.

Keywords: Pension fund, machine learning, investment performance, risk profiling, risk tolerance prediction, feature selection,Recursive Feature Selection, Boosting, Artificial Neural Networks, Xgboost, CART

JEL Classification Numbers: G11, G40.

1 Introduction

Assessing individual attitudes towards risk is closely related with understanding people's financial decision making process. Hence, finding reliable predictors of actual risk taking behavior is at the hearth of disciplines including psychology, economics and behavioral finance, among many others. In particular, determining the impact of personal and socioeconomic characteristics in predicting individuals' risk taking behavior may reveal many insights for policy markers too. After the global financial crash in 2008, regulators urged that advise given by financial institutions should be in line with customer risk preferences. The well known MiFID ("Markets in Financial

Instruments Directive") regulation has been developed on the basis of these regulatory concerns. This regulation developed in EU member countries has required the financial institutions to test the suitability of investment advise offered to retail customers. MiFID obliges investment firms to collect certain information about their clients. In some countries, survey questionnaires are developed directly by regulators while in others financial institutions develop their own . Even though the regulation is a good start for the retail investors' protection, the effectiveness of these regulatory efforts is also crucial to enable investors to make sound decisions . At present, there are very few studies reviewing the effectiveness of risk profiling regulation globally. In a global practice review paper by Brayman et al (2015) , it was claimed that 27.8% of all risk profiles had poorly worded questions and 75% of these surveys had a scoring model that had arbitrary weighting. More importantly, in this survey, researchers have stated that only 10% of the financial institutions who made risk profiling know how to do the back testing of risk profiles. Our aim in this paper is to fill this gap by conducting an empirical analysis on risk profiling by using a unique data obtained from a large pension fund in Turkey. We have investigated the actual individual risk taking behavior of this financial institution which consists of 81,563 individuals. The data is comprised of a risk survey questionnaire and some additional demographic attributes which are not used in the actual risk scoring. By using the recent machine learning tools, we analyzed the effectiveness of the risk survey questionnaire by relating them to the actual risk taking behavior of investors. Main target variable of the actual risk taking behavior variable is measured by the realized volatility of the portfolio of retail investors. We used the feature selection methods popular in machine learning to find out the most important determinants of actual risk taking behavior. The individual responses to the self risk assessment question in the risk survey appeared to be one of the most important variables behind the actual risk taking behavior. This finding is in line with the literature (see Gürdal et al (2017) and the references therein). We have observed that, self assessment question is the most important factor out of eight risk survey questions. A simple lottery question appeared to be the second important feature determining the actual risk taking attitude. Therefore, two out of eight risk survey questions are related to actual risk taking attitude in our analysis. We have also investigated relevance of other socioeconomic factors in determining the risk taking attitude. The second important variable in determining the risk taking behavior strikes as being the age of investors. There are similar findings in the literature as well. In addition, we also found that age may have a nonlinear effect on risk taking attitude. Up to a point, age and risk taking move in parallel, after a point, age and risk taking behavior begin to move in opposite directions. To model this observation we used the square of age as a proxy variable and this proxy factor turned out to be the other important feature in determination of risk taking. This finding is similar to Riley and Chow (1992) who found a parabolic relationship with risk aversion declining with increasing age until age 65. After this point an increase in risk aversion is observed. In addition, we also made a predictive analysis of various machine learning models against the risk score developed by the regulator. Ridge , Lasso Regression and Elastic net regression in addition to decision tree regression (RT)i random forest (RF) models, artificial neural network methods (ANN) and boosting methods such as Gboost, Xgboost and light boost methods are among the machine techniques we used. According to our findings based on sample data , between 12%-15% of the total variation in risk taking can be explained by surveys and demographic factors. Similar find-

ings was documented in (Foerster et al, (2017)). However, we found that many of the nonlinear machine learning models, particularly boosting models have a better out of sample prediction performance compared to the one used by the regulator. To the best of our knowledge, this is the first study employing machine learning techniques in this context. Our findings suggest that machine learning tools can be used in risk profiling industry. Feature selection and predictive analysis of machine learning tools may set a standard in testing and improving risk profiling practice. Validity of risk surveys and selecting the most important variables from the general demographics and socioeconomic variables can be done in a more systematic way. Organization of the paper is as follows. In the following section we will discuss the individual risk profiling in pension funds .In the third section, we will discuss the data and preliminary analysis on behavioral finance. In section four, we will discuss machine learning methods used in the feature selection. In section 5, we will discuss the machine learning methods for predicting the actual risk taking behavior. In the sixth section, we will discuss the empirical results. We will conclude in the final section.

2 Risk Profiling: Literature Review

There are several ways to measure retail investors' risk taking behaviour. One of them is to analyze the revealed risk preferences on the basis of incentivized field experiments with real money prizes. The other one is to analyze the real world risk taking decisions by focusing on the equity share of a portfolio or measuring the volatility of the portfolio . Finding the determinants of these actual risks taken in real world has many implications. Through this analysis, one can answer the explanatory power of risk surveys which would reveal great insights for the regulators. In addition, finding other demographic features that may be helpful for predicting risky behavior of individuals could be another arena to analyze the reliability risk profiling mechanism.

Wang and Hanna (1997) found that risk tolerance increases with age when other factors are controlled. These findings are generally in line with the literature . In this respect, Klement and Miranda (2012) point out that traditional industry approaches focus on socio-economic and demographic factors such as age, income, wealth, marital status, and gender in setting up individual portfolios. However, such factors explain only a fraction of the variation in portfolios.[1]

2.1 Risk Profiling: Pension Fund Industry in Turkey

Turkey's private pension system complements the state-provided social security retirement system. Legislated in 2001, it became fully operational in 2003 with six pension companies. The total amount of assets invested is around 20 billion USD in 2022, which is around 3% of GDP. It offers a defined contribution scheme where participation is voluntary. Individuals can invest

[1] Other factors may account for some of the unexplained variation in risk preferences such as genetic predisposition to financial risk (Barnea, Cronqvist, Siegel, 2010); the advisers' own characteristics ; whether the investor lives in a country with lower political stability or social cohesion (Wang, Rieger, Hens, 2016); or life experiences with financial risk (Malmendier Nagel, 2011; Ehrmann Tzamourani, 2012). In addition, investors' measured risk preferences may vary depending on context (Harrison List, 2004; Levitt List, 2007). As such, an investor might exhibit a certain set of risk preferences in a low stake game in a laboratory setting or while answering a questionnaire, and quite a different set in real life (Klement Miranda, 2012, p. 12).

in portfolios developed by asset management companies. [2] Participants are allowed to change their fund allocation six times in a given year. Unlike the funds in the US, Turkish investors cannot invest in individual stocks but in mutual funds on equities, bonds or even commodities. In 2017, the regulatory authority, Capital Markets Board of Turkey (CMB), required financial institutions to measure risk preferences of clients and provided financial recommendations based on their elicited risk attitudes consistent with the concept of suitability[3]. Since then, individuals investing in pension funds need to complete the survey developed the CMB . Accordingly, the risk classification score calculated by investors is an important base for them to allocate their pension funds. This procedure is in line with the Markets in Financial Instruments Directive (MIFID) used in the European Union. This regulation requires financial advisors to provide recommendations to investors, parallel to their investment objectives. There are various approaches in quantifying risk tolerance in the financial industry . In this specific survey designed by the CMB, there were seven questions to determine the risk tolerance of individuals. Response for each question has a certain weight which is used to calculate the final risk score for each individual. All the institutions in the pension system are obliged to apply this questionnaire as set by the regulator.

3 The Data

The data set used in this paper is provided by a large pension fund company including 81,563 customers during the period 2018-2022. In the data set, we have records of the responses to the questions in the risk questionnaire survey.XXXX The regulator has a fixed weight function which translates these responses into a risk score. The risk score of each individual is calculated by summing these responses. In addition, we have the information on gender, income, age, educational level, marital status, and contribution level. We also know whether the participant works in the private sector or public sector. The pension fund customers choose a portfolio of funds supplied by the company.XXXX In the pension system, individuals are not allowed to pick individual stocks or bonds for their portfolio. However, they can choose mutual funds based on equity, bond, commodities, etc. Each of these funds have different level of market risk. Actual riskiness level of each of these funds are measured by their respective monthly return volatility. The regulator classifies volatility of these funds between a scale of 1 to 7 where 1 being the least risky and 7 is the highest fund risk. For each investor, we know their actual funds and their rested XXXX realized volatility. Thus, we calculated actual volatility taken by each individual by using the weighted average of the funds' volatility investedXXXX. This is the actual risk taken by these individuals. Normally, the risk profiling score and the actual risk level taken should be suitable for each individual. For the suitability framework to be valid, individual investors' risk tolerance and the actual risk taken should be closely related. So the main question is how the estimated risk score and the actual risks taken are related. For a high linear association in this context, is risk tolerance and actual risk to happen the questions and the weights of the

[2]For detailed information on the evolution of the pension fund system in Turkey, see P ension Monitoring Center of Turkey. (https://www.egm.org.tr/homepage)

[3]The concept known as suitability which requires financial advisors to collect relevant information on the investment goals of the clients and their attitudes toward risk taking

risk score should be linked to the actual risk taken by the individual investorXXXXXXX. Our main research question is how to find the main factors behind the actual risk taken measure by annual volatility of the fund investedxxxxx.

3.1 Commenting on the data

Below we have some descriptive data on the risk profiling and funds invested. As can be seen from the graph, the mean age of the sample is around 40 years old. Participants seem to have a fairly good education with many College and Post College degrees.

Table 1: Descriptive Statistics: Demographics

	Married	College	Post-college	High school	Primary	Male	Private	Public	Sector-Other	Age
Sample size.	58347	45091	5595	24237	4733	55150	52853	9253	19457	81563
Mean	0.72	0.55	0.07	0.30	0.06	0.68	0.65	0.11	0.24	39.52
St.Dev	0.45	0.50	0.25	0.46	0.23	0.47	0.48	0.32	0.25	9.48
Min	0	0	0	0	0	0	0	0	0	18
Max	1	1	1	1	1	1	1	1	1	98

Table 2: Descriptive Statistics: Income, Age and Average Contribution

	Income	Age	Contribution
Mean	19,395	39.52	326.32
StD	26337.87	9.48	713
Min	0	18	0
Max	348,979	99	37,560

In our sample 68% of the participants are male and 32% are female. Majority of present demographic characteristics of the our sample.

Figure 1: Correlation Heat Map: Risk Survey Questions

Notes:

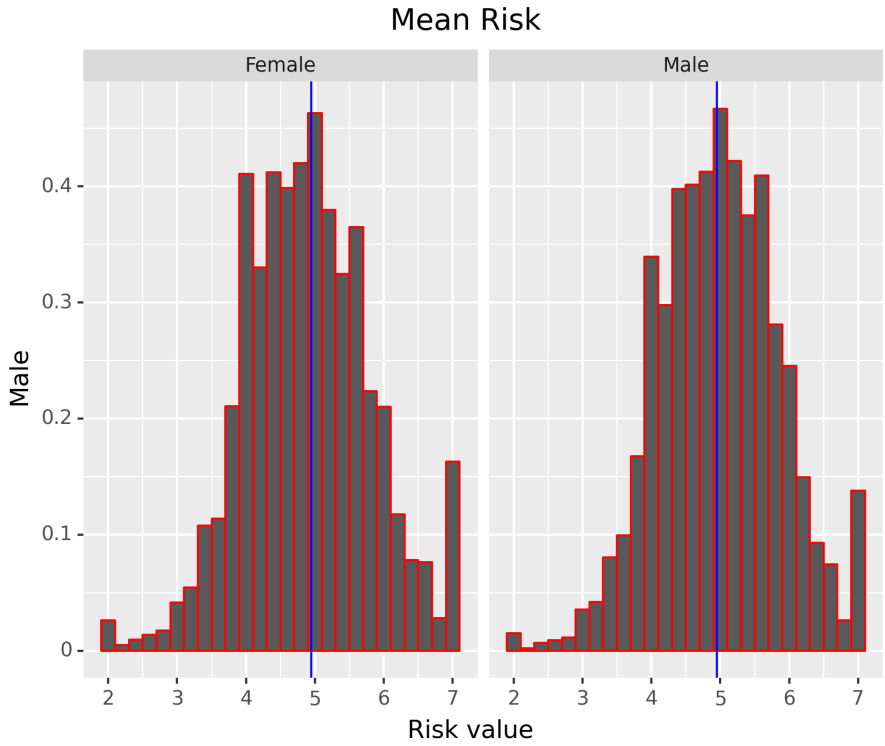

Figure 2: Portfolio Risk Level by Gender

Notes:

In the histograbelow, it seems that male investors tend to take more risk than that of malesXXXX. Risk value is in the graph refers to the actual risk taken by investors measured by volatility. As mentioned before, it lies between 0 and 7 and has a mean of 4.95.

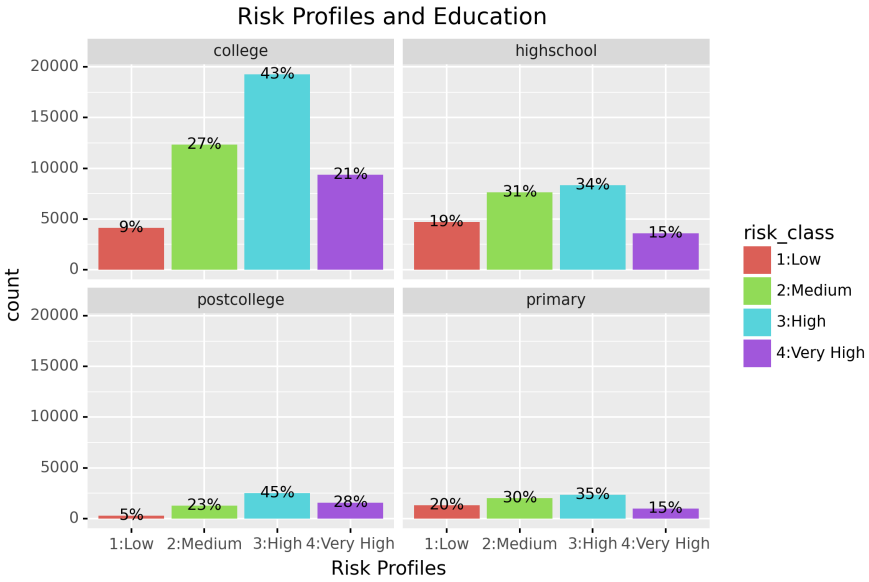

Figure 3: Risk Profiles by Educational Differences

Notes:

It is also important to see how the educational differences affect actual risk taking behavior. Surprisingly, there are less educational differences in actual risk taking behavior.

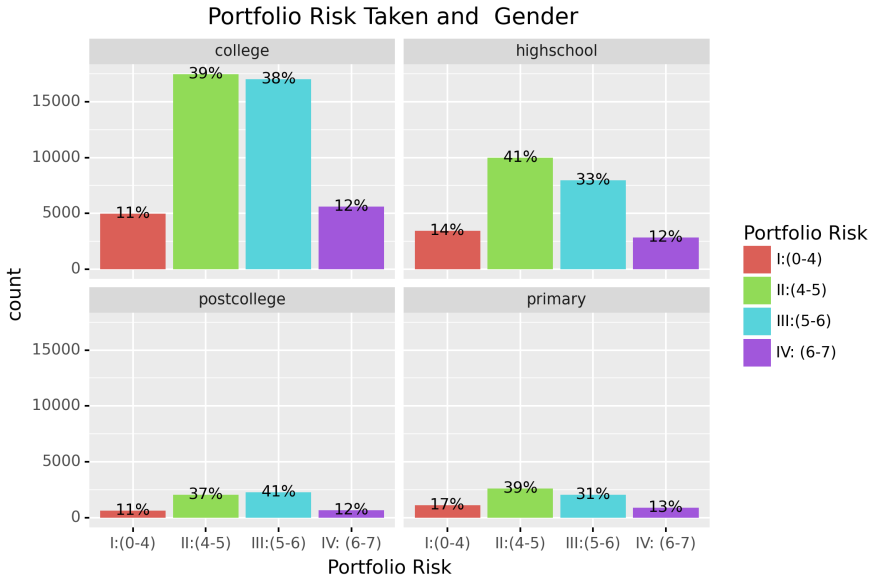

Figure 4: Risk Profiles by Educational Differences

Notes:

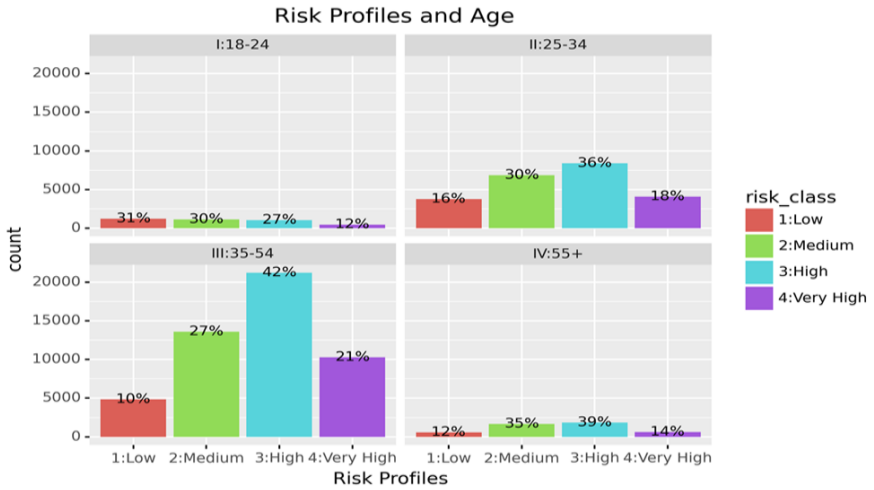

Figure 5: Risk Profiles by Aging

Notes:

As can be seen form the table and graph, age has a very important dimension both in risk profile and actual risk taking behavior. As age increases, people's risk taking behavior increases. Especiallly after the age of 35, investors take more risk. This trend works until mid 50's then risk taking behavior start to decline. This is very much in line with the pension fund research.

Table 3: Age Group and Risk Taking Behavior

Age Group	Portfolio Risk	Risk Score
18-24	4.82	20.68
25-34	4.99	24.28
35-54	4.97	26.00
55+	4.61	24.22

In general, we see a nonlinear effect of age on both the risk tolerance and actual risk taking behavior. Initially risk taking tendency increases with age. Then after a certain age it starts to decline.

Table 4: Descriptive Statistics on Income age and Contribution

	income	age	contribution
mean	19395.34	39.52151	326.320291
count	81563	81563	81563
std	26337.87	9.479269	713.0416145
min	0	18	0
255075max	348979	99	37560

Income variable data is not a dollar income, XXXXX(but) rather an estimate of income developed by the institution.the Pension fund makes a yearly income on the basis of some other factors. This is to avoid the possibility of some people not revealing their income beacuse of informality in some sectorsXXXX. Contribution in the fund varies on investors income level.

Table 5: Risk Profile Classification and Actual Risks Taken

Risk Profiles	Portfolio risk	Income
Low	4.412114	13290
Medium	4.689848	16905
High	5.078681	20629
V.High	5.428186	24644

As can be seen from the above graph, as income level increases, people's risk attitude seems to grow. The extent to which this observation is valid in statistical sense will be tested in the the following sections.

In the partbelow, we documented the risk survey responses and the total risk scores on the basis of responses. The average risk score is 25 but can get to 47.

Table 6: Results on Risk Profiling Questionnaire

	quest1	quest2	quest3	quest4	quest5	quest6	quest7	quest8	risk taken	risktotal
mean	2.97	2.34	3.10	2.58	2.80	2.55	3.30	1.76	4.95	25.16
std	1.04	1.00	1.15	0.98	0.76	0.87	1.12	0.43	0.89	8.47
min	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00	0.00
255075max	4.00	5.00	5.00	4.00	4.00	4.00	5.00	2.00	7.00	47.00

Since risk score developed by the regulator has fixed weights over the responses in the above table, we can see that responses and risk total scoreXXXX are moving in the same

direction. Question 8 is to test whether investors prefer to have interest rate bearing assets or not. The aim of this question is to distinguish whether investors have religious concerns over the choice of funds. In graph below, we can see how the risk tolerance and actual risk taking behavior are related for different educational level and gender. Almost in combinations of gender and educational background, risk tolerance and risk taking behavior seemed to be positively correlated. But to understand the sensitivity of this relationship more testing is required, which is left toXXXX the next section.

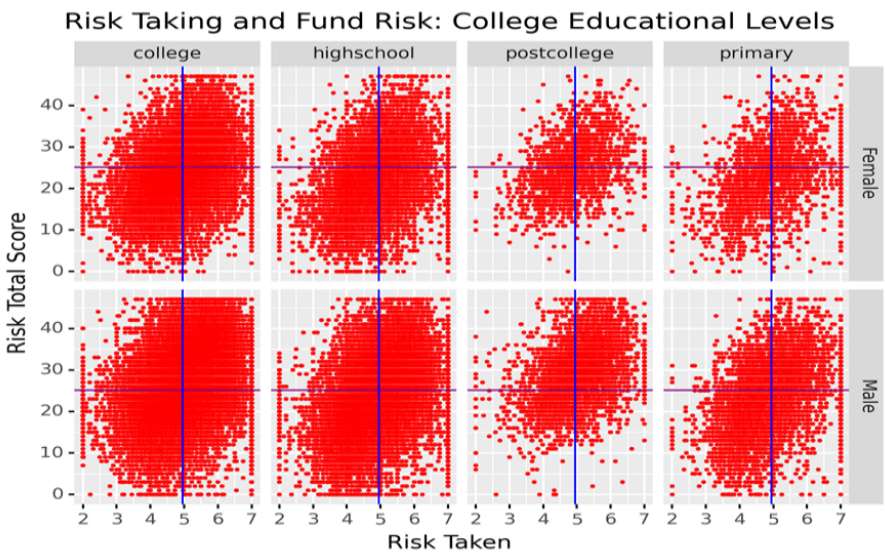

Figure 6: Risk Profiles by Aging

4 Machine Learning Techniques

Machine learning techniques have become very areaXXXX over the last decade and will be getting more trendy in finance and economics applications. See Hamima (et al, 2022) and Hoang and Wiegratzmakes(2021) for recent surveys on this area. Applications of behavioral finance and risk profiling are very rare. However, these methods in the finance industry are also very popular. For instance, most robo-advisory firms use these methods for allocating investors funds in a mathematical form. There two main application areas of machine learning, namely, predictive analysis and feature selection. Hence, we will first introduce some well known machine learning models and then introduce the feature selection methods in our context. Finally, we will complete the predictive analysis on our risk profiling data.

4.1 Machine learning methods

There are very different models that are applicable in the field of machine learning. The nature of the model differs in various ways but in the machine learning field, the most popular regression models can be given(are) as follows: Ridge,LASSO, Elastic Net Regression, Artificial Neural Networks,k-Nearest Neighbour Regression (kNNR), regression tree (RT) models, Random Forest

(RF). In addition Boosting models are very poular namely, Gradient Boosting Tree (GBT), Xtreme Gradient Boosting, Light Gradiendt boosting.

4.2 Ridge Regression

Ridge regression was a popular method (see Tibshirani et al (2021) for a detailed account of all these methods), where we need to dealXXXX with many explanatory variables or features in the machine learning jargon. Particularly, when there are many predictors potentially correlated, Ridge regression proposes a good model alternative. In this model, one can fit a model containing all k predictors using a technique that constrains or regularizes the coefficient estimates. The method in a way is equivalent to shrinking the irrelevant coefficient estimates towards zero. This method is computationalXXXX more compact and decreases the estimation variance. Suppose the standard linear regression is estimated by minimizing the sum of squares below.

$$RSS = \sum_{i}^{n}(y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2 \tag{1}$$

In order to reduce the variable dimension, k Ridge regression proposes the following penalty function.

$$\min \sum_{i}^{n}(y_i - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

where $\lambda \geq 0$ is a tuning parameter (or hyperparameter) which will be determined separately while minimizing RSS is standard but the additional *shrinkage constraint* needs to be minimizedXXXX. In this minimization process, βs need be close to 0 so it has the effect of shrinking penalty the estimates of β's towards zeroXXXXX. Therefore, variables that have marginal explanatory power will shrink to zero.

Choosing λ optimally is critical and is done via cross validation technique in statistics which is discussed in detail in Jame etl al (2021) and Hull(2021).

4.3 LASSO Regression

An alternative to Ridge regression is known as LASSO (Least Absolute Shrinkage and Selection Operator). It has a different penalty function which can be given as

$$\min \sum_{i}^{n}(y - \beta_0 - \beta_1 X_{1,i} - ... - \beta_k X_{k,i})^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{3}$$

λ will be determined separately by using cross validation techniques

$\lambda \sum_{j=1}^{k} |\beta_j|$ is the Lasso penalty,XXXX similar to Ridge, Lasso also tries to push the coefficients of variables that has marginal explanatory power to zero. Therefore, these methods can be useful when someone works with a very large potential explanatory variable set. Large variable dimension can be effecitvely reduced through this method . Similar to ridge, WhileXXXXX minimizing RSS is standard, but for Lasso penalty to be minimized we XXXXneed 's to be close

to 0. Similar to Ridge regression, λ which is known as tuning parameter can estimated via cross validation method.

4.4 Elastic Net Regression

Elastic Net regression is a relatively more recent method which can be considered as a mixture of Ridge and LASSO regressions. Elastic net regression method has an advantage over both of them . The model can be given as,

$$\min \sum_{i}^{n}(y - \beta_0 - \beta_1 X_{1,i} - ... - \beta_p X_{p,i})^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{4}$$

λ_1, λ_2 are tuning parameters

$\lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|$ is the Elastic Net penalty

A comparison of these 3 models can be given in the equation below.

$$\min RSS = \sum_{i}^{n}(y - \beta_0 - \beta_1 X_{1,i} - ... - \beta_p X_{k,i})^2 \tag{5}$$

$$\text{subject to:} \tag{6}$$

$$\sum_{j=1}^{k} |\beta_j| \leq s, \text{(Lasso)} \tag{7}$$

$$\sum_{j=1}^{k} \beta_j^2 \leq s, \text{(Ridge)} \tag{8}$$

$$(\sum_{j=1}^{k} |\beta_j| + \sum_{j=1}^{k} \beta_j^2) \leq s, \text{(Elastic Net)} \tag{9}$$

4.5 Regression Tree models

A popular method in ML is to use decision tree regression algorithms. Original idea of Classification and Regression Trees (CART) goes back to Brieman et al (1983).CART models can handle non-linearity better than standard linear models. Regression tree model is a non-parametric model which uses decision tree algorithm to estimate the target variables.Decision trees start with a question, such as, "Male or Female?" Tnen, you can ask a series of questions to determine an answer, such as, "Is it a long period swell?" Another positive aspect of these regression models is that they are easy to interpret. Regression Tree model involves selecting input variables (independent variables) and split points on those variables until a suitable tree is constructed. In order to estimate these methods it is suggested (see Hull(2021). We used Sklearn libraries in Pyhton and used Gini statistic to find out the dominant variables at each nodes.

4.6 Random Forest

RF method is an ensemle methods and an extension to CART model. It was originally proposed by Brieman(2001). Random Forest method selects the independent variable sets randomly

from the potential variable set. For each sample there is a corresponding independent variable. Therefore RF model uses a random subset sampled independently with the same distribution for RF model (see (Breiman, 2001). From these randomly created multiple trees final prediction will be based on the average of these samples. It has some advantages over standard RT models. It is more robust to noise than the standard tree models.

4.7 ANN

Artificial neural networks (ANN), is a technique inspired by the biological neural network. The simplest ANN structure is the one in which units distributes in two layers: An input layer and an output layer. With this set up it enables to handle complex relations in many disciplines. More details can be found in Hastie et al (2009).

4.8 k nearest neighbor

k nearest neighbor regression method, kNN, is a very old non-parametric method which is used both in academia and industry (see Hastie et al (2009)). In this method, the input consists of the k closest examples in a data set.This method generates forecasts according to the Euclidean distance computed between the points used for training and testing. Its simplicity and interpretability makes the method very popular.

4.9 Gradient Boosting

Boosting is an ensemble method where many models are combined to have a better prediction performance.The idea in boosting is to create strong learners from weak learners. The method developed by the computer scientists Kearns, M. and Valiant L. (1989) and statistician Friedman (1999).It is a sequential ensemble method where each new model tries to correct the errors of the previous models. In boosting, predictions are made sequentially, in each trial the previous error is being corrected. The technical advantage of boosting it can reduce both the bias and variance which is a very useful feature in prediction. The final predictions are made by combining the predictions of all the models. The method adapts the weights of weak learners so that it reaches to a better prediction.

4.10 Extreme Gradient Boosting:XGboost

XGBoost is a gradient boosting algorithm that combines weak learner models to create a strong model.The method developed by Tianqi and Guestrin (2016) uses a special type of tree called an extreme gradient boost tree. These trees better capture non-linearity in the data and handle large amounts of data efficiently. The key to XGBoost's success is its ability to efficiently train these trees while still maintaining high accuracy. This method is very efficient and has shown good performance in prediction in the practical forecasting exercises.

4.11 Light Gradient Boosting:lgboost

Ke et al (2017) claimed that despite the efficiency and scalability of XGboost, it may have some disadvantages when the feature dimension is high and data size is large. They modified the boosting algorithm via using a more efficient algorithm which is called as ligtGBM.

5 Important Factor Selection with Machine Learning

Variable or feature selection is a very important area of machine learning. In data analysis there are too many factors that might be related in explaining a target variable. Therefore the task becomes to select the most important features to explain a dependent variable. There are many different feature selection methods where in this paper we adapted recursive feature elimination (RFE). In our case we had 8 risk survey questions and 14 other factors. We wanted to select the most important variable subset of the whole potential variable set. By using the variable importance methods proposed by the machine learning methodology we estimated the Lasso regression and rank the absolute value of the coefficients. The variables that contribute most can be seen in the following graph. Responses to question 4 is also appeared to be important. All other demographic and survey questions appeared to be less important than others.

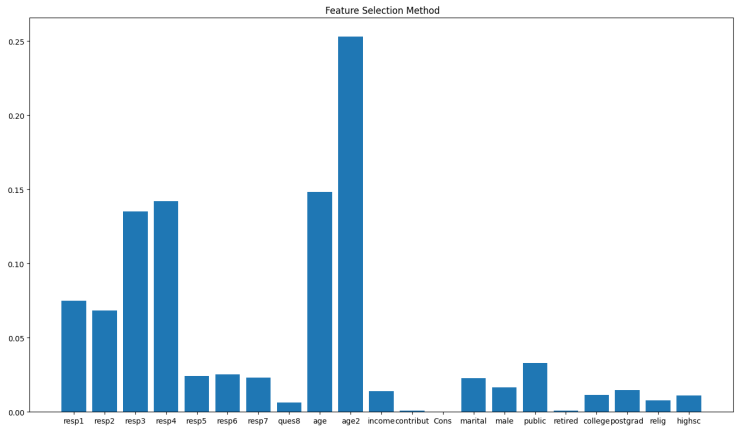

Figure 7: Variable Importance

As can be seen the self risk assessment question, age and square of age appeared to be the most important variables in explaining the variability of actual risk taking behavior.

5.1 Feature Selection further evidence

In below we tried another feature selection algorithm in the context of ML known as Recursive Feature Elimination. For each method we applied RFE method to find out which variable is important for determining the actual risk taking behavior. Main aim of the recursive feature elimination (RFE) is to select features by recursively targeting a smaller sets of independent variables.In this algorithm first a model is fit. Then the coefficients with the smallest value (most unimportant) is removed. Then the second unimportant feature is removed. Finally, we will reach the variable set with the most important features. One needs to stardardize the original features to be consistent is selection procedure. This algorithm fits a model and determines how significant features explain the variation in our data set. Once the feature importance has been determined, it then removes those less important features one at a time in each iteration. So for

each method we chose, we rank the features from 1 to last one. If feature ranks as top it means , for that particular method, that feature is the most important one. We divided the sample into training and test sets. Since we had a very long data we did not need to use cross validation. The results obtained for the

Table 7: Recursive Feature Selection (RFE)

	RT	ols	rf	gboost	XGBoost	LGBoost	ridge	Lasso	selection
q1	2	4	17	2	2	3	4	3	0
q2	4	3	16	4	3	2	3	4	0
q3	1	1	1	1	1	4	1	1	7
q4	1	2	1	1	1	6	2	2	4
q5	5	5	14	5	5	8	5	5	0
q6	14	7	13	9	8	7	7	7	0
q7	13	8	12	7	9	5	8	9	0
male	12	10	11	12	14	11	10	10	0
age	3	1	10	1	1	1	1	1	6
age sq	1	1	9	3	17	16	1	1	4
income	6	11	15	6	7	1	11	11	1
q8	11	16	8	11	6	9	16	17	0
contribut	10	15	7	10	11	1	15	16	1
college	9	13	6	14	13	13	13	14	0
graduate	8	12	5	17	15	17	12	12	0
high.sch	7	14	4	15	16	14	14	13	0
retired	15	17	3	16	12	15	17	15	0
public	16	6	2	8	4	10	6	6	0
marital	17	9	1	13	10	12	9	8	1

Supporting the previous tests, questions 3,4 and age variables are among the most popular variables in determining actual decision taking process.

As can be seen from the above Table, almost all models suggest that the most important feature in predicting actual risk is question 3 which is about a self assessment risk question. This finding supports Gürdal et al (2017) and many others simple and direct questions are helpful in predicting risk. Age on the other hand the other important variable. In 6 out of different machine learning models age variable appears at the top ranked variable. This means in predicting risk age should be included. It is important to note that age also appears to affect risk taking behavior in a nonlinear way. We used age square as a proxy for potential non-linearity in the risk prediction. It turns out that 4 out of 9 models age square is the top ranked variable. Finally, a simple lottery question, question 4, reported in the Appendix is another important variable. In most of the models we used in the

6 Machine Learning Predictions

In this section we used the 85,163 individual data with the following definitions. The data collected by the pension fund between 2018-2022.

In this section we will present the models used in predicting the risk behavior. In comparison we used first the validation test approach. In the experiment design we have chosen 80 percent

of the data for training and 20 for the testing. So the training Set was 62,250 individuals and Test Set was 19,263. The data mainly randomly chosen. Then each of the models Mean Square Error is given as below. Since the data is large it is fairly satisfactory in terms of comparison. For all of the analysis we used Sklearn library python. We both used a very powerful PC with 64 Mbyte Ram and Google Cloud.

Table 8: Risk Predictions with ML models:Validation Set

method	MSE
ols	0.693436443
ridge	0.693436276
lasso	0.732010824
elnet	0.818645422
knn-	0.729221353
cart	0.713179496
ann	0.691101723
rforest	0.699166282
gboost	0.684526574
xgboost	0.688383517
lgboost	0.68374018
svr	0.7005274
regulation	0.712918599

6.1 Comments on findings: Validation Approach

We used standard OLS method and the risk score of the authority as two benchmarks. We used the actual risk score and run a linear regression to predict the actual risk taken by the individuals. As can be seen in the above table many of the nonlinear models can beat the risk score calculated by the regulator. We studied mainly nonlinear ML models' performance namely, k-nearest neigbour (knn) Artificial Neural Networks (ANN) ,lightgboost, Xgboost, gradientboost and Regression Trees for Machine Learning (RT). Most of the models seem to do better than OLS. Mainly the boosting methods, ann do produce better result than that of OLS. Therefore, hidden nonlinearities in the risk choice allows the alternative models to fare better than OLS. It is also important to add more of the other demographics to improve the actual risks taken.

6.1.1 Model Comparison by Cross Validation

There are basically two approaches in backtesting the predictive performance of machine learning models namely validation set approach and cross validation. In validation set approach we divide the sample into two which are training set and test set. We do train our models in the training set and use these estimates to predict out sample. We used 20 % of the whole sample for the out of sample prediction. The rest will be used for training. On the basis of these predictions we calculate the Mean Square Error for these models.

The second approach is called cross validation. In this method, called k-fold CV, the training set is split into k smaller sets . The following procedure is followed for each of the k "folds":

A model is trained using of the folds as training data the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such

as accuracy). The performance measure reported by k-fold cross-validation is then the average of the values computed. This approach is in general computationally demanding , but use the valuable data.In this section we have used alternative ML models to predict financial risk taken by individuals. We have trained these models through k-fold cross validation with k=10. We divided the sample into 10 and trained each model for these different 10 subsamples. We could see not only the mean but standard deviation of MSE here. The average of these folds is the average MSE of these test sets. As can be seen in below a fair amount models beat linear model. Particularly Boosting base models were more successful than that of OLS. Some of the methods such as SVR was computationally so demanding we left out this method for the below analysis.

Table 9: Model Comparison with cross validation

Model	mean(R2)	st.dev(R2)
regulator's score	0.131422734	0.009529904
ols	0.154649867	0.009187395
knn	0.134879228	0.009725188
cart	0.130687787	0.007821762
ann	0.151987501	0.009554904
rf	0.147245536	0.008629616
gboost	0.166079702	0.009042466
xgboost	0.16263462	0.008893901
lgboost	0.167172251	0.0086789
ridge	0.154649909	0.009186948
lasso	0.153763655	0.008791215
elasticnet	0.154645359	0.009152423

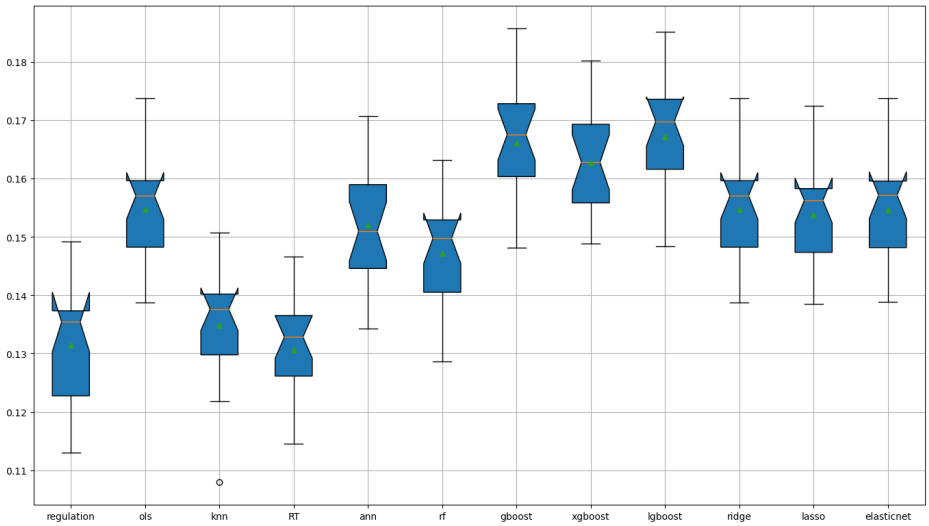

Figure 8: Model Comparison R2

Table 10: Model Comparison: Only with selected features R2

Model	mean(R2)	st.dev(R2)
regulator's score	0.1314	0.0095
ols	0.1336	0.0089
knn	0.1168	0.0100
cart	0.1320	0.0079
ann	0.1357	0.0085
rf	0.1382	0.0079
gboost	0.1431	0.0080
xgboost	0.1421	0.0084
lgboost	0.1442	0.0079
ridge	0.1336	0.0089
lasso	0.1309	0.0084
elasticnet	0.1336	0.0089

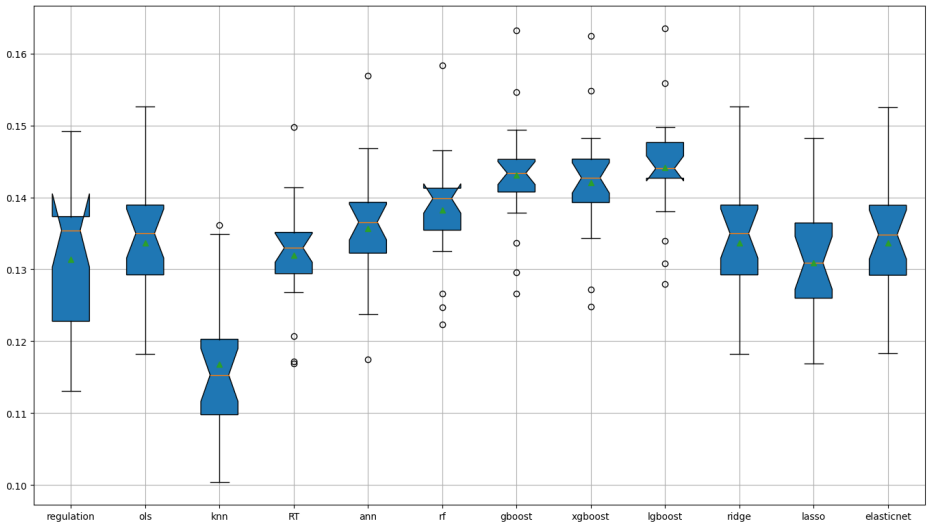

Figure 9: Model Comparison: Best selected features and benchmark

7 Conclusion

With a unique pension data we analyzed the individual risk taking behavior of investors with a large data set. We found that direct and comprehensible questions work better in understanding the actual risk behavior. More comprehensive questions might be difficult to grasp for a general risk surveys. In line with the findings in the literature, we found out that age is the most critical element in risk taking in pensions. We also found out the more religous people tend to take less financial risk than average. We have found that among many variables age and age square are the most critical variables in determining actual risk taking behavior. In addition, we have

found that machine learning techniques can produce better risk forecast precision. Particularly boosting methods have a better prediction behavior than linear methods. We believe the risk surveys and their relationship between actual risk taking decisions require more attention both in academia and in the financial regulation field.

8 References

[1] Agnew, J., Balduzzi, P., & Sunden, A. (2003). Portfolio choice and trading in a large 401(k) plan.*American Economic Review*, 193-225.

[2] Agnew, J. (2010). Pension Participant Behavior. In H. Kent Baker and John Nofsinger (Eds.), *Behavioral Finance Investors, Corporations, and Markets* (pp. 577-594). Hoboken, NJ: John Wiley & Sons, Inc.

[3] Ameriks, J., & Zeldes, S. P. (2004). How do household portfolio shares vary with age? (Working Paper). Columbia University.

[4] Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, 55(2), 773-806.

[5] Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, 116(1), 261-292.

[6] Benartzi, S., & Thaler, R. H. (2007). Heuristics and biases in retirement savings behavior. *Journal of Economic Perpectives*, 21(3), 81-104.

[7] Bucher-Koenen T., Ziegelmeyer M, Once Burned, Twice Shy? Financial Literacy and Wealth Losses during the Financial Crisis, *Review of Finance*, 18-6, 2215–2246

[8] Calvet, L. E., Campbell, J. Y., & Sodini, P. (2007). Down or out: Assessing the welfare cost of household investment mistakes. *Journal of Political Economy*,115, 707-747.

[9] Calvet, L. E., Campbell, J. Y., & Sodini, P. (2009). Fight or flight? Portfolio rebalancing by individual investors. *The Quarterly Journal of Economics*, 124(1), 301-348.

[10] Choi, J. J., Laibson, D., Madrian, D., & Metrick, A. (2002). Defined contribution pensions: Plan rules, participant choices, and the path of least resistance. In James M. Poterba (Ed.), *Tax Policy and the Economy 16* (pp. 68-113). Cambridge, MA: MIT Press.

[11] Choi, J. J. (2015). *Contributions to defined contribution pension plans* (NBER Working Paper No. 21467). Retrieved from http://www.nber.org/papers/w21467

[12] Christiansen, C., Joensen, J. S., & Rangvid, J. (2015). Understanding the effects of marriage and divorce on financial investments: The role of background risk sharing. *Economic Inquiry*, 53(1), 431-447.

[13] Clark, R., Lusardi, A., & Mitchell, O. S. (2015). Financial knowledge and 401 (k) investment performance: A case study. *Journal of Pension Economics and Finance.*

[14] Elkind, D., Kaminski, K., Lo, A,W., Siah,K.W.,Wong,C.H. (2021), When Do Investors Freak Out?: Machine Learning Predictions of Panic Selling SSRN,*https://papers.ssrn.com/sol3/papers.cfm?abstractic*

[15] French, K. R., (2008). Presidential address: The cost of active investing. *The Journal of Finance*, 63(4), 1537-1573.

[16] Graham, J. R., Harvey, C. R., & Huang, H. (2009). Investor competence, trading frequency and home bias. *Management Science*, 55(7), 1094-1106.

[17] Iyengar, S., & Kamenica, E. (2010). Choice proliferation, simplicity seeking, and asset allocation. *Journal of Public Economics*, 94(7), 530-539.

[18] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-292.

[21] Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246, 160-173.

[22] Kuzubaş, T.U., Saltoğlu, B., Sert, A., Yüksel, A., (2019) "Performance Evaluation of the Turkish Pension Fund System", *Journal of Capital Markets Studies*, 3, 18-33

[23] Kimball, M. S. and Shumway, T. (2010) Investor sophistication and the home bias, diversification, and employer stock puzzles. University of Michigan Working paper.

[24 Madrian, B., & Shea, D. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *Quarterly Journal of Economics*, 116, 1149-1188.

[25] Mitchell, O. S., Mottola, G. R., Utkus, S. P., & Yamaguchi, T. (2006). *The inattentive participant: Portfolio trading behavior in 401 (k) plans* (Pension Research Council Working Paper 2006-5).

[26] Mitchell, O. S., Todd, P., & Bravo, D. (2007). Learning from the Chilean experience: The determinants of pension switching (Working Paper). University of Chile, Department of Economics.

[27] Odean, T. (1999). Do investors trade too much? *American Economic Review.* 89(5), 1278-1298.

[28] OECD. (2021). Pension Funds in Figures. retrieved from https://www.oecd.org/finance/private-pensions/Pension-Funds-in-Figures-2021.pdf

[29] Rooij, M., Lusardi, A., & Alessie, R. (2007). Financial literacy and stock market participation. *Journal of Financial Economics*, 101(2), 449-472.

[30] Talpsepp T., Liivamägi K. & Vaarmets, T. (2020) Academic abilities, education and performance in the stock market *Journal of Banking and Finance* 117,1-25.

[31] Vaarmets, T., Liivamägi K., Talpsepp T. "From academic abilities to occupation: What drives stock market participation?" *Emerging Markets Review*, 39, 83-100

[32] Yamaguchi, T., Mitchell, O. S., Mottola, G. R., & Utkus, S. P. (2007). "Winners and losers: 401 (k) trading and portfolio performance" Working Paper. Michigan Retirement Research Centre, University of Michigan.

[33] Ahmed, S., Muneer M. , Alshater, Anis El Ammari, Helmi Hammami (2022). "Artificial intelligence and machine learning in finance: A bibliometric review", *Research in International Business and Finance*, 61, 101646.

[34] P. Pudil, J. Novovicova, J. Kittler (1994), Floating search methods in feature selection, *Pattern Recognition Lett.* 15 (1994) 1119–1125.

[35] Breiman L, Friedman JH, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth Brooks, Pacific Grove

[37] Breiman L (2001) Random forests. *Mach Learn* 45(1): 5–32

[38] Marinelli and Mazzoli, 2011,

https://www.tandfonline.com/doi/abs/10.1080/13669877.2014.910678

[39] Rui Mata, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. (2018). Risk Preference: A View from Psychology, Journal of Economic Perspectives, 32, 2—Spring

2018—Pages 155–172

[40] Flawed Self-Assessment: Implications for Health, Education, and the Workplace Risk Preference: A View from Psychology

[41] Levon Barseghyan, Francesca Molinari, Ted O'Donoghue, and Joshua C. Teitelbaum (2018), Estimating Risk Preferences in the Field, Journal of Economic Literature, 56(2), 501–564

[42] Clark, R., Lusardi, A., & Mitchell, O. S. (2015). Financial knowledge and 401 (k) investment performance: A case study. *Journal of Pension Economics and Finance*. doi: 10.1017/S1474747215000384

[43]Dhar, R., & Zu, NR. H. (2007), Up Close and Personal: Investor Sophistication and the Disposition Effect*Management Science*, 52(5), 726-740

[44] David Dunning, Chip Heath, Jerry M. Suls (2004), https://journals.sagepub.com/doi/full/10.1111/j.1529-1006.2004.00018.x$_i$1

[45] Kearns, M. and Valiant L. (1989). Cryptographic limitations on learning Boolean formulae and finite automata. Symposium on Theory of Computing. Vol. 21. ACM. pp. 433–444. doi:10.1145/73007.73049. ISBN 978-0897913072. S2CID 536357.

[46] Spiliotis, E. and Makridakis, S., Semenoglou. A, Assimakopoulos. V (2022), Comparison of statistical and machine learning methods for daily SKU demand forecasting *Operational Research*, 10.1007/s12351-020-00605-2

https://www.wealthbriefing.com/html/article.php?id=195859.Y1lqynZBy70

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., "Gene selection for cancer classification using support vector machines", *Mach. Learn*, 46(1-3), 389–422, 2002.

Brayman,S., Finke,M., Bessner,E. Spry S. , Grable,J., Humber, G. , Clement R. (2015), Current Practices for Risk Profiling in Canada And Review of Global Best Practices https://www.osc.ca/sites/default/file 02/iap$_2$0151112$_risk - profiling - report.pdf$

[47] Tianqi, C. and Guestrin C. (2016), XGBoost: A Scalable Tree Boosting System https://arxiv.org/pdf/1603.02 Guolin, Key, Qi, M., Finley, T., Wangl, T., Ma, W. , Ye Qi,Tie-Yan Liu (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree, https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848 Paper.pdf

9 Appendix:Risk Profile Questionnaire

Q1. I plan to stay in the pension system for the period specified below.

(a) Less than 3 years

(b) 3 to 6 years

(c) 6 to 10 years

(d) More than 10 years

Q2. How do you evaluate your knowledge about financial transactions?

(a) I have no knowledge.

(b) I have very little knowledge.

(c) I have a reasonable level of knowledge and understand the difference between investment instruments.

(d) I am highly knowledgeable and understand the factors that may affect my investments.

(e) I am an experienced and expert investor.

Q3. What do you think about return and risk taking?

(a) My primary goal is to ensure that my savings do not lose value by taking the lowest possible risk, rather than the return expectation.

(b) I prefer to earn a small regular return and take low risk in this direction.

(c) I would like to take a moderate risk to get a reasonable return.

(d) My primary goal is to grow my savings, capital losses and fluctuations that may occur in the short term do not worry me.

(e) I want to get the highest return in the long run. I can take high risk for this.

Q4. Assume that you have 4 different investment options in a one-year period where you obtain positive or negative returns. Which of the following funds would you prefer for your 100,000TL (Turkish Lira) savings?

(a) **Fund A**: 3% return (3000TL return)

(b) **Fund B**: Equal chance to gain 10% return or 3% loss. (10000TL gain or 3000TL loss)

(c) **Fund C**: Equal chance to gain 25% return or 10% loss. (25000TL gain or 10000TL loss)

(d) **Fund D**: Equal chance to gain 50% return or 20% loss. (50000TL gain or 20000TL loss)

Q5. Your retirement savings of 10,000TL have lost value to 9,000TL. In this case:

(a) I panic and sell all my funds. I try not to take risks.

(b) I get nervous. I switch to different and safer funds for some of my savings.

(c) I wait patiently before making any changes. I think my savings will gain value again in the long run.

(d) I give more weight to funds that cause a decrease in my savings when their prices have decreased. I turn low prices into an opportunity.

Q6. Which of the following summarizes your or your family's financial situation?

(a) I have no/very little savings and a large amount of debt.

(b) I have some savings and debts that I pay regularly.

(c) I can save regularly and my debt is at a level that can be paid off at any time.

(d) I have sufficient amount of savings and I have no debt.

Q7. When we think about retirement, do you think you will have enough income?

(a) I do not have a pension from Social Security Administration, I do not expect any income other than private pension system.

(b) I do not have any other security other than the pension that my spouse will receive from Social Security Administration.

(c) I do not have any other security other than the pension I will receive from Social Security Administration.

(d) In addition to the pension I will receive from Social Security Administration, I expect additional income from my real estate investments.

(e) I have other financial investments or participated in other pension systems.

Q8. Is it important to you whether your investment includes interest or not?

1. Yes, I prefer it to be interest free.

2. No, it is not important.