# Benchmarking Regression Algorithms for Income Prediction Modeling

Azamat Kibekbaev*, Ekrem Duman
Industrial Engineering Department
Özyeğin University
Istanbul, Turkey
E-mail: kibekbaev.azamat@ozu.edu.tr , ekrem.duman@ozyegin.edu.tr

***Abstract*— This paper aims to predict incomes of customers for banks. In this large-scale income prediction benchmarking paper, we study the performance of various state-of-the-art regression algorithms (e.g. ordinary least squares regression, beta regression, robust regression, ridge regression, MARS, ANN, LS-SVM and CART, as well as two-stage models which combine multiple techniques) applied to five real-life datasets. A total of 16 techniques are compared using 10 different performance measures such as R2, hit rate and preciseness etc. It is found that the traditional linear regression results perform comparable to more sophisticated non-linear and two-stage models.**

***Keywords— Regulation; Income prediction; Regression techniques; Performance measures.***

## I. INTRODUCTION

The credit card market has exploded over the years with increasing interest to the computerization of society. It is one of the most widely accepted ways to pay around the world. Credit cards revolutionized consumer spending habits, changed the face of business. In today's market, consumers are demanding more personal attention. They expect companies to understand their needs and offer products and services that meet those needs almost instantaneously. As a result, today's lenders compete by offering complex and customized products to specific individuals. In doing so, they take into account consumers' tastes for different cards, how valuable these consumers might be as customers, and other offers these individuals might receive from competitors. For this reason, in today's economy, credit cards represent an important part of household, business and global activity.

Credit limit is the most important feature of a credit card. It is an amount of credit that a financial institution extends to a client or in a simple words amount that can be spend in total with credit card. The credit card limit of an individual will be decided by a combination of factors, where lenders are likely to look closely at the spending power and income, as this allows them to determine how much the individual can afford to borrow. Individual consumer income information enables the banks to validate application data, target new prospects, and segment existing customers. Income can be understood as the summation of all earnings (wages, salaries, profits, interests payments, rents etc.) in a given period of time (generally a month). Income is a crucial demographic element that is used at a wide variety of customer touch points. Therefore, it is very important to have an income prediction for existing and potential customers. However, accurate indicators of income are difficult to collect but essential for companies that want to create high-quality revenue budgets, especially in an uncertain economic environment with changing government policies. Increased competition have made banks and financial institutions to search for new ways to minimize the denial of credit to creditworthy customers and to keep out fraudulent ones as far as possible. Thereby enabling them to offer the right product at the right stage with good relations. Accordingly, they need rapid and correct decision making process in order to maintain high borrowing power. The advantage for financial institutions is their individual segmentation of customer's data that facilitates analysis to categorize the optimum combination of outcome of risks and assets overtime. Under these conditions, the subjective judgment of decision makers is a crucial factor in making accurate forecasts to provide solutions that not only assess the creditworthiness, but also keep the per-unit processing cost low, while reducing turnaround time for the customers.

Our aim and motivation in this paper is to predict incomes of customers for Turkish banks, in relation to some new banking regulations in Turkey. Turkey's banking regulator (BDDK) has announced new regulations which brought about a series of stricter rules on the use of credit cards. Major rule was to launch a "single limit" for credit cards, where consumer's credit card limit could not exceed four times the amount of his/her monthly income to offset the uncontrolled use of loans and credit cards in the domestic market and to decrease household debt levels. All of these limits will be applicable to all banks in Turkey. So that the sum of the limits from different banks cannot exceed this "single limit". Income prediction will let us determine credit card limit for each individual customer and allow us to control credit card holders not to exceed their upper limits. Therefore, it is crucial to have models that estimate income as accurately as possible. Also, in near future similar limitations will also be applied to personal loans.

The rest of the paper is structured as follows: Section 2 presents a brief literature survey. Section 3 describes different regression methods and the performance measures used in this study. Section 4 briefly explains datasets and experimental set-ups in regression techniques. Results of the estimation are

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.162

discussed in Section 5. Finally Section 6 concludes the paper and provides directions for future research.

## II. LITERATURE SURVEY

In the literature there are only a few empirical studies on income prediction. This is justified by the fact that it is highly difficult to get exact information about individuals income, wealth and their characteristics. Here we concentrate on individual income models and predictions but the modeling is often limited by lack of adequate data. For example, as Carrier and Sand argued that employers do not easily volunteer to give salary data [6]. Bone and Mitchell showed that obtaining more appropriate data and good model for retirement income elements can lead to better estimation [2]. Thereby, Using U.S. sample data, Dominitz presented income prediction by comparing datasets from 1993 and 1994 [10]. He finds that income expectations were optimistic on average. In contrast, Das and van Soest examined data from the Dutch Socio-Economic Panel between 1984 and 1989 where they find that income expectations were too pessimistic on average [9]. By examining 18 years of monthly data from the Michigan Survey of Consumer Attitudes and Behavior, Souleles wanted to show that most variables appear to have been biased and inefficient when predicting income [16].

Based on literature survey, most income prediction studies were about determination of future incomes for college students. They study the effect of a student's college GPA, major, and standardized test scores in order to see what is most influential on future income. Thomas and Smart stated that college performance leads to higher earnings after graduation [15, 17]. Chia and Miller used data from the University of Melbourne in Australia in order to study the effect of college performance in future income [8]. They find that GPA is playing major role in starting salaries.

Regression analysis is a statistical technique for investigating or estimating the relationship among variables. It is used when you want to predict a relationship between a dependent variable and one or more independent variables. Regression analysis can be of two types: nonlinear and linear. In this work we will use some type of regression techniques. This is why, we wanted to provide some articles from the literature related to income estimation by regression models.

Carlos determined an accurate estimated gross margin of the farms using regression models and types of neural network [5]. Results from ANN models, have provided the most accurate gross margin predictions rather than regression models. Chen and Yang proposed quasi-stepwise regression variable selection method based on validation of least absolute deviation regression to forecast rural household net income [7]. According to study the models constructed by the quasi-stepwise regression variable selection method and the least absolute deviation estimation method have lower errors, higher validation and can be applied to many other forecasting issues. Also, benchmarking works were done by Loterman et al. (2012) to estimate loss given default (LGD) using various regression techniques for 6 real-life datasets to model and predict LGD [13]. According to their observation, non-linear techniques such as LSSVM, ANN and MARS perform significantly better than linear ones. Also, their work showed that the variance in LGD remains poorly explained because of the resulting techniques have limited explanatory power.

This paper is the one of the first attempts to predict incomes to regulate credit limit of bank customers. Since little is known about income prediction in credit cards, we want to fill the void of academic literature with our large-scale income prediction benchmarking study using 16 different regression techniques as shown in Table 1 and five real-life income datasets from Turkish banks to estimate and regulate income for credit limits.

## III. REGRESSION TECHNIQUES AND PERFORMANCE MEASURES

Regression techniques are separated into one-stage and two-stage models, where one-stage methods are also divided to linear and non-linear ones. In Linear models our developed techniques are named as originals and transformed, as sometimes we need to transform our dependent variable to more closely meet the assumptions of a statistical inference procedure. For that reason, original techniques (OLS, RiR and RoR) are used directly to datasets without any replacement or change in dependent variables, while transformed ones (BR, B-OLS and BC-OLS) used to meet the OLS normality assumption or to improve the normality of the dependent variable. Linear and non-linear techniques were part of one-stage models; two-stage models are combinations of one-stage non-linear techniques with OLS to add predictive power for techniques.

TABLE I. REGRESSION TECHNIQUES

| Techniques | Descriptions |
|---|---|
| **Linear** | |
| **OLS** | Ordinary least squares regression is a generalized linear modelling technique that minimizes the sum of squared residuals. |
| **B-OLS** | This model fits the Beta distribution to the dependent variable to transform that variable before estimating our OLS model. |
| **BR** | Model is commonly used by practitioners to model variables that assume values in the standard unit interval (0, 1). It is based on the assumption that the dependent variable is beta-distributed and that its mean is related to a set of repressors through a linear predictor with unknown coefficients and a link function. |
| **BC-OLS** | Box and Cox have proposed a family of transformations that can be used with non- |

| | |
|---|---|
| | negative responses and which includes as special cases all the transformations in common use, including reciprocals, logarithms and square roots [3]. This model also improves the normality of the dependent variable before OLS. |
| **RiR** | Technique for analyzing multiple regression data that suffers from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. |
| **RoR** | Is an alternative to least squares regression when data are contaminated with outliers and it can also be used for the purpose of removing influential observations from the least-squares fit. |
| **Non-Linear** | |
| **CART** | CART was introduced by Breiman et al. in 1984 [4]. A recursive partitioning method, builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification) in order to improve the fit as best as possible. |
| **MARS** | Method is presented for flexible regression modeling of high dimensional data. MARS is a stepwise procedure for the automatic selection of base functions from observed data and it is a non-parametric regression technique for solving regression-type problems, with the main purpose to predict the values of a continuous dependent or outcome variable from a set of independent or predictor variables. |
| **LSSVM** | Method close to SVM formulation but solves linear system instead of QP problem, which related to regularization networks and Gaussian processes but additionally emphasize and exploit primal-dual interpretations from optimization theory [18]. |
| **ANN** | Nonlinear statistical data modeling tool where the complex relationships between inputs and outputs are modeled. In this study we used the popular multilayer perceptron (MLP). A Multilayer Perceptron (MLP) is a neural network that is trained using back propagation. |
| **M5P** | M5' method [14,19], this is a decision tree with linear regression functions at the leaves. It can be used to predict a numeric target (class) attribute. It produces a piecewise linear fit to the target. |
| **OLS + Non-Linear** | |
| **OLS+RT, OLS+ANN etc.** | These two-stage algorithms integrate predictive power of non-linear models and OLS. In the first stage, OLS with linear model is built. In the second stage, non-linear models estimate the residuals of linear model. Finally, estimated residuals from non-linear models are added to OLS estimation. |

The performance of prediction models can be assessed using a variety of different methods and metrics. Performance measure helps us decide which algorithm is better or worse than another and measures how well a data mining algorithm is performing on a given dataset. We have two different metric categories as calibration and discrimination. Calibration is a measure of how well predicted values agrees with actual observed values, while discrimination is the ability of the model to correctly separate the subjects into different groups or ability to provide an ordinal ranking of the dependent variable. Also, one of the main properties of performance measure is ranges that it can take. Ranges of RMSE and MAE are from 0 to infinity with 0 being the perfect score. AUC and AOC can range between 0 and 1. In AUC higher values show how good the model is, however in AOC lower values are better. In general, $R^2$ or coefficient of determination takes a number on the scale between 0 and 1, where 1 is perfect positive correlation and 0 no correlation. Discrimination metrics such as Pearson, Kendall and Spearman ranges from -1 to 1. Correlation is 1 if the agreement between the two rankings is perfect, -1 if the disagreement is perfect and 0 if there is no correlation. Hit rate and preciseness are in calibration category in evaluating model performance. Both of them are ranges from 0 to 1, higher values show how good the model makes predictions (ratio of the number of predicted hits to the number of observed).

TABLE II. PERFORMANCE MEASURE

| Metric | Descriptions |
|---|---|
| **RMSE** | Root-mean-square error measures the differences between values predicted by a model and the values actually observed. |
| **MAE** | Mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcome. |
| **AUC** | The area under the curve (AUC) of a receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. Which is used here to plot how good the model is at |

| | |
|---|---|
| | distinguishing values that are higher than average from those that are lower than average. |
| **AOC** | REC curves in order to select a good threshold value, so that only residuals greater than that value are considered as errors [1]. The REC curves facilitate visual comparison of regression functions and they are qualitatively invariant to choices of error metrics and scaling of the residual. |
| **RSquare** | R-square is the square of the correlation between the response values and the predicted response values. It is also called the square of the multiple correlation coefficients and the coefficient of multiple determinations. |
| **Pearson's r** | Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by *r*. |
| **Spearman's ρ** | Spearman's rho is a statistical measure of the strength of a monotonic relationship between paired data. |
| **Kendall's τ** | Kendall's tau provides a distribution free test of independence and a measure of the strength of dependence between two variables. |
| **Hit rate** | Hit rate is a term used to describe the success rate of an effort or it is a ratio or a proportion. This rate specifically compares the number of times an initiative was successful against the number of times it was attempted. In our study, we use hit rate to measure how successful are our algorithms to make an approximate prediction of incomes. In other words, if the predicted income is within ±α percent of actual income, we will consider it as a hit. |
| **Preciseness** | Preciseness provides a quality or state of being very accurate. It measures success of algorithm; does our predicted income is within acceptable interval or it overestimates or underestimates the real income. In other words, level of success in shooting accurately at targets in a given range. |

## IV. Datasets And Preprocessings

The number of dataset entries are demonstrated in Table 3, where we have five real-life datasets from Turkish banks. We can see that there are 10,000 observations in each dataset except Dataset 3. The number of available input variables ranges from 10 to 15 and all of them are continuous. Each dataset is randomly shuffled and separated into 30% test set and 70% train set. The training data is used to build the model by pairing the inputs with the expected output. The test data used to estimate how good models have been trained and to estimate model properties or to assess the prediction performance of the models.

After some data preprocessing, variable selection procedure was implemented in order to eliminate redundant or irrelevant features. The objectives of attribute selection are: providing more cost-effective predictors and improving the performance of regression model. In our study, variables in all datasets were similar and we used stepwise selection and in some cases expert opinion to decrease the number of attributes. Stepwise selection is a combination of backward elimination and forward selection. It is a method that allows moves in direction, dropping or adding variables at various steps.

TABLE III. DATASET CHARACTERISTICS

| **Datasets** | Inputs | Total size | Training | Test |
|---|---|---|---|---|
| **Dataset 1** | 10 | 10000 | 7000 | 3000 |
| **Dataset 2** | 10 | 10000 | 7000 | 3000 |
| **Dataset 3** | 15 | 13018 | 9100 | 3918 |
| **Dataset 4** | 13 | 10000 | 7000 | 3000 |
| **Dataset 5** | 12 | 10000 | 7000 | 3000 |

## V. Results And Discussion

The performances of the resulting techniques are measured on the test sets according to the ten performance metrics that was listed in Table 2. Final model performance results for all 16 regression algorithms obtained from five real-life datasets are shown in Table 4.

Considering final results, linear models built by OLS, RiR and RoR showed similar performances between each other in all ten metrics. In contrast, transformed linear models such as BR, B-OLS and BC-OLS perform much worse than original linear techniques. Despite the fact that these approaches are specifically designed to cope with a violation of the OLS normality assumption.

According to our results, non-linear techniques (i.e. MARS, ANN, LS-SVM and M5P) outperformed the linear models in all datasets except CART. CART demonstrated worse performance in accordance with other non-linear models. Being better than linear techniques explains that income (dependent variable) and independent variables in the datasets are non-linearly correlated.

In this research to compare our model performances we concentrated on $R^2$, Hit rate and Preciseness which are defined in Table 2. According to this, the highest performance values for given metrics were in non-linear and two-stage models, where two-stage models significantly outperform most of the non-linear techniques in income prediction.

By looking at Table 4 one can notice that the results for hit rate and preciseness with $R^2$ are not parallel in datasets. So that, while a technique is bad in $R^2$, it can perform good in hit rate or preciseness. Therefore, $R^2$ values may not always be explanatory and it depends on business objectives or needs.

Good $R^2$ does not indicate whether a regression model is adequate for a given dataset. One can have a low $R^2$ value for a good model, or a high $R^2$ value for a model that does not fit the data. For this reason, using additional performance measures will be better for making decisions.

Eventually, we determined the relative ranks for each algorithm on all five datasets. Then, by taking the average of those ranks using Friedman's test [11], we calculated some overall rank as displayed in Table 4. It is a nonparametric statistical test performed to test the null hypothesis that all regression techniques perform alike, based on their rankings for a chosen performance metric. Once Friedman's test rejected the null hypothesis, we proceeded with a Hommel's [12] post-hoc test method in order to find the concrete pairwise comparisons which produce rank differences with 95% confidence level. According to demonstrated result of these tests the difference between models were small. Thus, Table 4 gives a concrete suggestion to practitioners, so that, instead of implementing and testing all regression algorithms, one may select the top five best performing techniques (OLS+M5P, OLS+LS-SVM, OLS+MARS, M5P and MARS) for implementation and comparison.

TABLE IV. AVERAGE RANKING FOR ALGORITHMS

| Rank | R^2 | | Hit rate 15% | | Preciseness (ACC) | | Total Rank | |
|---|---|---|---|---|---|---|---|---|
| 1 | OLS+M5P | 2,6 | OLS+M5P | 2 | OLS+M5P | 3,6 | OLS+M5P | 2,7 |
| 2 | OLS+LSSVM | 2,8 | OLS+LSSVM | 2,4 | OLS+LSSVM | 4 | OLS+LSSVM | 3,1 |
| 3 | OLS+MARS | 3,8 | M5P | 3 | OLS+MARS | 4,2 | OLS+MARS | 4 |
| 4 | M5P | 5,2 | MARS | 4 | MARS | 4,6 | M5P | 4,3 |
| 5 | MARS | 5,2 | OLS+MARS | 4 | M5P | 4,8 | MARS | 4,6 |
| 6 | OLS | 6,2 | LSSVM | 4,8 | LSSVM | 5,4 | LSSVM | 6,1 |
| 7 | OLS+ANN | 7,8 | OLS | 6,6 | Robust | 7,2 | OLS | 6,9 |
| 8 | Robust | 8,2 | OLS+ANN | 7,4 | OLS | 8 | Robust | 7,7 |
| 9 | LSSVM | 8,2 | Robust | 7,8 | Ridge | 8,2 | OLS+ANN | 7,8 |
| 10 | ANN | 8,2 | Ridge | 8,6 | OLS+ANN | 8,2 | Ridge | 8,5 |
| 11 | Ridge | 8,8 | ANN | 8,6 | ANN | 9,4 | ANN | 8,7 |
| 12 | OLS+CART | 11,8 | OLS+CART | 10,4 | OLS+CART | 10,4 | OLS+CART | 10,8 |
| 13 | CART | 12,4 | Beta/OLS | 12,6 | CART | 13,4 | CART | 12,8 |
| 14 | Beta/OLS | 14 | CART | 12,6 | Beta/OLS | 14 | Beta/OLS | 13,5 |
| 15 | BoxCox | 15,2 | BoxCox | 14,2 | BoxCox | 14,6 | BoxCox | 14,6 |
| 16 | BetaReg | 15,6 | BetaReg | 15 | BetaReg | 15,6 | BetaReg | 15,4 |

## VI. CONCLUSION

In this large-scale income prediction benchmarking study, we evaluated 16 different regression techniques on five real-life datasets obtained from Turkish banks. The regression performance was assessed by 10 different performance metrics (MAE, RMSE, ROC, Spearman and Kendall etc.). It was found that the non-linear and two-stage techniques except CART yield very good performances in terms of all performance measures. However, it has to be noted that simple linear regression also had very good performance, which clearly indicate that for most datasets, income can also be calculated by the traditional linear regression. The experiments also indicated that many regression techniques yield performances which are quite competitive with each other (as MARS, M5P, ANN and LSSVM). Starting from the findings of this study, several interesting topics for future research can be identified. One of this promising avenues for future research would be to repeat the same experiment for customer groups of different characteristics (such as self-employed ones, retired ones etc.). Another interesting topic would be to make classification work by considering dependent variable as nominal, for example, denoting incomes as high, medium, and low. Also, we intend to explore more methods and datasets from different countries to see the differences of country-based income prediction.

## VII. ACKNOWLEDGMENT

We would like to thank anonymous Turkish banks who made this study possible by providing us with in-house data on income of their customers.

## VIII. REFERENCES

[1] Bi, J., Bennet, K. P. (2003). Regression error characteristic curves. In Twentieth international conference on machine learning.

[2] Bone C. & Mitchell O. (1997). Building Better retirement Income Models. North American Actuarial Journal, 1, 1-12.

[3] Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, 26, 211–252.

[4] Breiman L., Friedman J.H., Olshen R.A. and Stone C. J. (1984). Classification and Regression Trees (2nd Ed.). Pacific Grove, CA; Wadsworth.

[5] Carlos R. G., Mercedes T., and César H., (2010). Income prediction in the agrarian sector using product unit neural networks. European Journal of Operational Research. Vol 204, 355–365.

[6] Carrier J. & Shand K. (1998). New Salary Functions for Pension Valuations. North American Actuarial Journal, 3, 18-26.

[7] Chen Q., and Yang C., (2008). Quasi-Stepwise Regression Variable Selection and its Application in Rural Household Net Income Forecasting. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. 28(11): 16–22.

[8] Chia, Grace and Miller, Paul W., (2008). "Tertiary Performance, Field of Study, and Graduate Starting Salaries. The Aus tralian Economic Review. 41 (1), pp. 15-31 .

[9] Das, Marcel, and Arthur van Soest, (1999). " A Panel Data Model for Subjective Information on Household Income Growth," Journal of Economic Behavior & Organization 40: 409-26.

[10] Dominitz, Jeff, (1998). "Earnings Expectations, Revisions, and Realizations," The Review of Economics and Statistics 80(3): 374-388.

[11] Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. Ann Math Statist, 11(1), 86-92.

[12] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika, 75, 383–386.
[13] Loterman, G., Brown, I., Martens, D., Mues, C., Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting, 28, 161-170.
[14] Quinlan, J.R. (1992). Learning with continuous classes. Proceedings of 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, pp. 343-348.
[15] Smart, John C., (1988). "College influences on graduates' income levels." Research in Higher Education. September , 29 (1), pp. 41 -59.
[16] Souleles, Nicholas S., (2003). "Consumer Sentiment: Its Rationality and Usefulness in Forecasting Expenditure– Evidence from the Michigan Micro Data," Journal of Money, Credit, and Banking, forthcoming (NBER working paper 8410).
[17] Thomas, Scott L., (2000). "DefelTed Costs and Economic Returns to College Major, Quality, and Performance." Research in Higher Education, 41 (3), pp. 28 1-313.
[18] Vapnik V. and Lerner A., (1963). "Pattern Recognition Using Generalizd Portait Method," Automation and Remote Control, Vol. 24, pp. 774-780.
[19] Wang, Y., Witten, I.H. (1997). Induction of model trees for predicting continuous classes. Proceedings European Conference on Machine Learning, Prague, pp. 128-137.