sciendo International Journal of Mathematics and Computer in Engineering 2(1) (2024) 117–124

International Journal of Mathematics and Computer in Engineering

https://sciendo.com/journal/IJMCE

Original Study

Revenue forecast models using hybrid intelligent methods

Gizem Topaloğlu[1], Tolga Ahmet Kalaycı[1], Kaan Pekel[1], Mehmet Fatih Akay[2] †

[1]Trendyol, Data Science, İstanbul, Türkiye

[2]Department of Computer Engineering, Cukurova University, Adana, 01330, Türkiye

Communicated by Haci Mehmet Baskonus; Received: 14.06.2023; Accepted: 22.07.2023; Online: 31.10.2023

Abstract

The aim of this study is to forecast the revenue of a seller taking part in an online e-commerce marketplace by using hybrid intelligent methods to help the seller build a solid financial plan. For this purpose, three different approaches are applied in order to accurately forecast the revenue. In the first approach, after applying simple preprocessing steps on the dataset, forecast models are developed with Random Forest (RF). In the second approach, Isolation Forest (IF) is used to detect outliers on the dataset, and minimum Redundancy Maximum Relevance (mRMR) is utilized to select the features that affect the quality of revenue forecast, correctly. In the last approach, a feature selection process is performed first and then the Density-Based Spatial Clustering and Application with Noise (DBSCAN) is used to cluster the dataset. After these processes are carried out, forecast models are developed with RF. The dataset used includes the daily revenue of a seller with several other features. Mean Absolute Percent Error (MAPE) is used for evaluating the performance of the forecast models.

Keywords: Revenue forecasting, machine learning, hybrid methods, marketplace.
AMS 2020 codes: 68Txx; 68T07; 68T20; 68T01; 68T09.

1 Introduction

Strategic planning, organizing, directing, and control of all financial activities within a company or institution constitute financial management. Along with being crucial to fiscal management, it also entails applying management principles to an organization's financial assets. Financial planning and budgeting ensure that the company has enough liquidity. A company that does financial planning has created a financial map. It assesses the causes of deviations from the target dates. Based on income and expense statements, financial planning identifies financial weaknesses and strengths as well as needs. Companies can expand their range of products or services and pursue new opportunities for methods that will increase their productivity and profitability with financial planning that allows them to make decisions that keep the company's value at the highest level. As a result, it can compare its position in the industry to that of other companies. It is necessary for good financial planning to understand the future values of some financial components and to manage the related processes accordingly. Making accurate forecasts for the relevant components is critical. In the last few years, numerous methods have been used for revenue forecasting. In [1], Chen et al. predicted the direction of one-year earning

†Corresponding author.
Email address: mfakay@cu.edu.tr

sciendo
ISSN 2956-7068
doi:10.2478/ijmce-2024-0009

changes using machine learning methods and high-dimensional financial data. The models that outperformed used Logistic Regression (LR), small sets of accounting variables, and professional analyst estimates. In [2], Chung et al. investigated how different machine learning models performed in revenue estimation for local methods and compared the performance of various machine learning algorithms in revenue estimation. The findings revealed that traditional statistical methods outperformed machine learning algorithms in predicting the property tax revenue of K-Nearest Neighbors (KNN). Via [3], Kureljusic et al. investigated and analyzed prediction models generated by machine learning algorithms using publicly available data. When compared to financial analysts, machine learning algorithms provided more accurate revenue forecasts. In [4], Lin et al. presented Generalized Additive Models (GAMs) and machine learning models based on Artificial Neural Networks (ANNs) that were developed to predict the optimal revenues of an integrated power generation and storage system. Based on optimized solutions from the Conventional Hydroelectric Power and Environmental Resource System (CHEERS) model, predictive equations and models were developed. Model validation prediction errors of GAMs and machine learning models were less than 5%; regression equations in machine learning models performed better. In [5], Mousa et al. proposed using earnings per share as a performance metric to predict corporate financial performance, employing three supervised machine learning methods: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Random Forest (RF). They also used a sample of 63 publicly traded banks from eight emerging markets between 2008 and 2017. The study concluded that the best prediction model was created with the RF, and evidence about the accuracy and performance of the presented models was discovered. Models had less than 5% error; regression equations in machine learning models performed better. In [6], authors explained the detail of machine learning applications for financial forecasting. Machine learning appears to be ideally adapted to enhance forecasting, analysis and planning by substantially automating the information extraction process from massive datasets. The simulation carried out in this paper demonstrated the potential of machine learning for forecasting and planning. Additionally, as the number of data points rose, the development of forecasting and planning was investigated. In [7], authors contrasted the Capital Asset Pricing Model's (CAPM) performance with machine learning algorithms and methodologies for predicting the price of financial assets, as well as implementations of machine learning algorithms on High-Performance Computing infrastructures. On out-of-sample test data, machine learning models beat CAPM after being trained on time series data. In [8], authors used the machine learning and deep learning approaches to provide a reliable forecast and total corporate profits in the US economy. The primary tool used to present this prediction method was the Rapid Miner software. Drawing on these predictions and based on economic theory, this article explored the implications of assumptions made to date regarding the relations between the working class and the elite. In [9], authors used the deep learning methods in the financial industry. The maturity of technology was also evaluated in different areas of the financial sector. Deep learning was not yet the most used technology in financial industry, but research showed that some problems in this area required deep learning features. In [10], a hybrid approach was proposed by combining Simple Linear Regression (SLR) and RF to increase the estimation accuracy. In addition, it was recommended to use effective floor space for training SLR and RF models. The hybrid approach of the proposed estimation methodology proved effective in reducing the risk of Building Information Modeling labor cost estimation. In [11], authors showed that large scale financial time series experiments could produce more accurate forecasts than those made by professional financial analysts. In [12], the cash flows of accounts receivable have been estimated using methods applicable to companies with a large number of customers and transactions. Before moving to neural networks with MultiLayer Perceptron, forecasting techniques such as Autoregressive Integrated Moving Average and Prophet and Long-Short Term Memory (LSTM) networks that have not been used for cash flows until now were discussed. In [13], authors showed the hierarchy of importance of financial factors of institutions and tried to make prediction with deep learning/machine learning models. A comparison was made between the recommended Extreme Gradient Boosting and deep LSTM models to aid the research. In [14], authors examined revenue projections made by financial professionals and identified factors that affect forecast accuracy. The benefit of revenue estimations was projected using a model that was thus created. Analysts who perform worse in predicting sales are more

likely to give up than analysts who perform better. The study helped academic researchers and investors in their understanding of the factors that influence income estimates. In [15], the feasibility of Chinese privately traded companies as a distressing example was analyzed using statistical methods and a prediction model based on Support Vector Machines (SVM) was developed. The grid search technique, which used 10-fold cross-validation, was used to find the best parameter value of the core function of the SVM. The SVM model outperformed traditional statistical methods and back propagation neural networks.

The rest of this paper is organized as follows. In section 2, dataset generation is given briefly. In section 3, methodology is presented by considering isolation forest, minimum redundancy maximum melevance, density based spatial clustering and application with noise and random forest. In section 4, results and discussion is reported in detail. In section 5, the conclusion is introduced by giving the results of this paper.

2 Dataset Generation

The dataset includes 1826 rows of data on a daily basis between January 1st, 2017 and December 31st, 2021 and includes the revenue of a seller along with some other features. The attributes in the dataset and their explanations are given in Table 1.

3 Methodology

3.1 Isolation forest

IF is used to find outliers and abnormalities. By calculating how far a data point is off the average, it isolates outliers rather than modeling the typical points. The alternative approach of IF that explicitly isolates outliers using binary trees shows a new potential of a faster anomaly detector that directly targets abnormalities without profiling all the typical cases. The linear time complexity of the method, the small constant requirements, and minimal memory usage make it successful when dealing with massive volumes of data [16]. The statistics of the revenue are given in Table 2 and the same obtained after removing the extreme values are given in Table 3.

3.2 Minimum redundancy maximum melevance

The feature selection has a big impact on how well estimate algorithms perform. The feature selection method consists of identifying and selecting the most advantageous aspects of the dataset. This technique has a substantial impact on the machine learning model performance. Unnecessary features can increase the model's error rate when the test dataset's input data considerably differs from the training dataset, lengthen the model's training time, and induce over fitting, which makes the model successful in the training dataset but fail in the test dataset. For these reasons, the mRMR algorithm [17] was used for feature selection in order to improve the performance of the models to be developed for revenue forecasting. The features selected using the mRMR are; Weekend, Year, Weekday, EURO_open, Total_covid_cases, USD_open, EURO_open, New_covid_tests, EURO_max, USD_min, EURO_close, Total_covid_deaths, USD_max, USD_close, Total_covid_tests.

3.3 Density-based spatial clustering and application with noise

Using the unsupervised learning technique known as clustering analysis, the data points are separated into numerous distinct bunches or groups, with the aim of ensuring that the characteristics of the data points within the same group are similar and those of the data points within different groups are somewhat different. In this study, the DBSCAN [18] clustering method was applied. The Density-Based Clustering concept is an unsupervised learning method that identifies distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. Within the scope of the study, the data set was divided into 5 clusters using the DBSCAN

method. Component values of the data sets obtained after clustering are given in Table 4.

Table 1 Attributes found in the dataset.

Attribute Name	Definition
Year	Year
Month	Mont
Day	Day of the month
Weekday	Weekday
Weekend	Weekend
USD_close	Closing value of the USD
USD_open	Opening value of the USD
USD_max	Maximum value of the USD
USD_min	Minimum value of the USD
EUR_close	Closing value of the EUR
EUR_open	Opening value of the EUR
EUR_max	Maximum value of the EUR
EUR_min	Minimum value of the EUR
Bist100_close	Closing value of the Bist100
Bist100_open	Opening value of the Bist100
Bist100_max	Maximum value of the Bist100
Bist100_min	Minimum value of the Bist100
Bist100_capacity	Capacity value of the Bist100
Total_covid_cases	Total number of people who caught in Covid
New_covid_cases	Number of new Covid cases
Total_deaths	Total number of people who died from Covid
New_deaths	Number of new deaths from Covid
New_covid_tests	Number of new tests
Total_covid_tests	Total number of tests
New_covid_vaccinations	Number of new vaccines
Total_covid_vaccinations	Total number of vaccines

Table 2 Statistics of revenue.

Statistics Name	Values
Number_of_lines	1826
Minimum	0
Maximum	4758634.92
Mean	1353609.99

Table 3 Statistics of revenue after applying IF method.

Statistics Name	Values
Number_of_lines	665
Minimum	0
Maximum	4654363.55
Mean	1242224.18

Table 4 Statistics of revenue after using DBSCAN.

Cluster	Number of lines	Minimum	Maximum	Mean
Cluster 1	399	1051826.12	1234772.39	1431150.39
Cluster 2	41	734682.64	762463.19	789182.86
Cluster 3	190	862860.11	956598.34	1044740.19
Cluster 4	115	94800.86	147999.73	202836.34
Cluster 5	168	1433005.72	1497854.28	1568416.04

3.4 Random forest

The RF is a supervised learning technique for regression. RF generates numerous decision trees throughout the training phase and averages the classes in order to anticipate all trees. From the training set, RF randomly selects k data points. A decision tree containing k data points is consequently produced. After selecting the desired number of trees, each tree built predicts the y-value for each data point [19]. The values sought and found by grid search for RF hyperparameters are given in Table 5.

Table 5 Hyperparameter values of RF.

Hyperparameter Range	Model Hyperparameter Values
"min_samples_leaf":[3,4,5,6]	min_samples_leaf: 5
"min_samples_split":[3,4,5,6]	min_samples_split: 5
"n_estimators":[50, 100, 200]	n_estimators:100

4 Results and discussion

Three different approaches were applied in order to accurately forecast the revenue. In the first approach, after applying only simple preprocessing steps to the dataset, forecast models were developed with RF. In the second approach, IF was used to detect outliers on the dataset, and the mRMR feature selection algorithm was utilized to correctly select the features that affect the quality of revenue forecast. In the last approach, the feature selection process was performed first and then the DBSCAN was used to cluster the dataset. After these processes were carried out, forecast models were developed with RF. The dataset used includes the daily revenue of a seller among several other features and covers the time period from January 1st, 2017 to December

31st, 2021. Grid search was used to obtain the best values of the hyperparameters of the RF method. The performance of the developed models was evaluated using MAPE. MAPE's of the models developed with three different approaches are shown in Table 6.

Table 6 Hyperparameter values of RF.

Approach	MAPE (%)
First approach	24.20
Second Approach	16.95
Third Approach/Cluster 1	7.69
Third Approach/Cluster 2	1.90
Third Approach/Cluster 3	4.66
Third Approach/Cluster 4	17.42
Third Approach/Cluster 5	2.36

The average of the MAPE values obtained for the five clusters in the third approach is 6.80.

– When the prediction models developed with the first approach and the second approach are compared, it has been determined that the MAPE obtained with the second approach is 7.25 % lower. When this result is evaluated, it has been determined that developing the prediction model by removing the outliers from the data set and applying the feature selection algorithm gives more successful results.

–When the prediction models developed with the first approach and the third approach are compared, it has been determined that the MAPE obtained with the third approach is 17.40 % lower.

– When the prediction models developed with the third approach and the second approach are compared, it has been determined that the result obtained with the third approach is 10.15 % lower.

– As a result of the comparisons, it has been determined that when the forecast model is developed for each cluster by dividing the data set into clusters, more successful results can be obtained in the revenue forecast.

– Results show that the IF outlier detection algorithm improves the performance of the models.

5 Conclusion

Corporate financial planning teams can leverage future forecasts created with machine learning-based algorithms by drawing on historical data to identify potential opportunities that will impact the future and growth of the business plan according to predicted consumer behavior, and stay ahead of the competition. In this way, they can quickly respond to changes and have the ability to make flexible but sound financial decisions through simulations of different scenarios. Three different approaches were thus applied to forecast revenue in this study. In the first approach forecast models were developed with RF. In the second approach, IF algorithm and mRMR algorithm were used. In the last approach, the feature selection process was performed first and then the DBSCAN was used to cluster the dataset. After these processes were carried out, forecast models with RF were developed. The results show that the lowest MAPE value was obtained with the third approach. It is finally observed that the average MAPE of the third approach is 17.40% lower than that of the first approach, and 10.15% lower than that of the second approach.

6 Declarations

6.1 Conflict of interest:

The authors hereby declare that there is no conflict of interests regarding the publication of this paper.

6.2 Funding:

Not applicable.

6.3 Author's contribution:

G.T.-Conceptualization, Data Curation, Methodology, Formal Analysis, Writing-Original Draft, Writing Review Editing. T.A.K.-Software, Data Curation, Validation. K.P.-Formal Analysis, Data Curation. M.F.A.-Supervisior. All authors read and approved the final submitted version of this manuscript.

6.4 Acknowledgement:

Not applicable.

6.5 Data availability statement:

All data that support the findings of this study are included within the article.

6.6 Using of AI tools:

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

References

[1] Chen X., Cho T., Dou Y., Lev B., Predicting future earnings changes using machine learning and detailed financial data, Journal of Accounting Research, 60(2), 467–515, 2022.
[2] Chung I.H., Williams D.W., Do M.R., For better or worse? Revenue forecasting with machine learning approaches, Public Performance and Management Review, 45(5), 1133–1154, 2022.
[3] Kureljusic M., Reisch L., Revenue forecasting for European capital market-oriented firms: a comparative prediction study between financial analysts and machine learning models, Corporate Ownership and Control, 19(2), 159–178, 2022.
[4] Lin Y., Li B., Moiser T.M., Griffel L.M., Mahalik M.R., Kwon J., Alam S.M.S., Revenue prediction for integrated renewable energy and energy storage system using machine learning techniques, Journal of Energy Storage, 50, 104123, 2022.
[5] Mousa G.A., Elamir E.A.H., Hussainey K., Using machine learning methods to predict financial performance: Does disclosure tone matter?, International Journal of Disclosure and Governance, 19, 93–112, 2022.
[6] Wasserbacher H., Spindler M., Machine learning for financial forecasting, planning and analysis: Recent developments and pitfalls, arXiv:2107.04851, 2021.
[7] Ndikum P., Machine learning algorithms for financial asset price forecasting, arXiv:2004.01504, 2020.
[8] Allen J., Giacoman K., Analytical approaches to macroeconomic forecasting: a study of profits through machine learning and deep learning, https://digital.wpi.edu/pdfviewer/nc580q60t, Accessed: January 1, 2020.
[9] Piispanen N., Sundqvist R., Vuotila R., Matilainen V., Emerging Technology Adoption and Use, Chapter: Applications of Deep Learning in Finance, ISBN: 978-952-03-1572-6, 1–181, 2020.
[10] Huang C.H., Hsieh S.H., Predicting BIM labor cost with random forest and simple linear regression, Automation in Construction, 118(103280), 1–16, 2020.
[11] Papadimitriou A., Patel U., Kim L., Bang G., Nematzadeh A., Liu X., A multi-faceted approach to large scale financial forecasting, Proceedings of the First ACM International Conference on AI in Finance, 15-16 October 2020, New York, USA, 5, 1–8, 2020.
[12] Weytjens H., Lohmann E., Kleinsteuber M., Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet, Electronic Commerce Research, 21, 371–391, 2021.
[13] Rayhan M., Sultana S., Majid A., Financial factors analysis for acquisition premium and anticipation using extreme

gradient boosting and deep recurrent neural network, B.Sc. Thesis, Brac University, Dhaka/Bangladesh, 1–58, 2019.
[14] Lorenz T., Homburg C., Determinants of analysts' revenue forecast accuracy, Review of Quantitative Finance and Accounting, 51, 389–431, 2018.
[15] Ding Y., Song X., Zen Y., Forecasting financial condition of Chinese listed companies based on support vector machine, Expert Systems with Applications, 34(4), 3081–3089, 2008.
[16] Cheng Z., Zou C., Dong J., Outlier detection using isolation forest and local outlier factor, Proceedings of the Conference on Research in Adaptive and Convergent Systems, 24-27 September 2019, Chongqing, China, 161–168, 2019.
[17] Ding C., Peng H., Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology, 03(02), 185–205, 2005.
[18] Kumar K.M., Reddy A.R.M., A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method, Pattern Recognition, 58, 39–48, 2016.
[19] Xu R., Improvements to random forest methodology, Ph.D. Thesis, Iowa State University, Lowa/USA, 1–87, 2013.