Store Sales Prediction Using Gradient Boosting Model

Jaeyoung Choi[1]

Library and Information Science,
Sungkyunkwan University, Seoul, 03063 Korea
cjengy@g.skku.edu

Abstract. Recently due to the development in machine learning and deep learning, applications of these technologies have been widely utilized daily and industrially. Implementations of machine learning on finance data have been in interest as well. Herein, we apply machine learning algorithms onto store sales data and present future applications for Fintech industries. We consider various missing data processing methods and utilize gradient boosting machine learning algorithms: XGBoost, LightGBM, CatBoost to predict the future sales for individual stores. As a result, we found that using simple median imputation with a XGBoost model has the best accuracy. By employing the proposed method, stores which have low credibility but have high probability to compensate for repayment can benefit by receiving assistance beforehand, while Fintech enterprises can benefit by offering financial instruments to these stores.

Keywords: Machine learning, Sales Prediction, XGBoost, LightGBM, CatBoost.

1. Introduction

Nowadays the rapid developments regarding machine learning and deep learning has led to various applications. The utilization of these technologies not only include information retrieval, computer vision, human computer interactions, robotics and natural language processing, but also approaches in fintech, where finance and technology is integrated.

There have been diverse appliances where machine learning is implemented for financial areas. Recent studies can be divided into two fields: Classification and Regression. For a project in classification tasks, is the study [11] which predicts the delinquency of members in a credit card company by credit card usage and deposit history. By the implementation of SVM(Support Vector Machine) which separates classes through hyper planes, a binary classification in whether or not the members will have credit delinquency in the future is implemented. With additional applications of ensemble modeling, this study concluded that ensemble methods had a better accuracy rate rather than singular models. Since the data is severely imbalanced, using methods such as LightGBM, which is a gradient boosting machine that deals imbalanced data in a better way, will be an advanced improvement. Another classification task[16], uses the UCI Credit Card Data set from Kaggle[7], to compare the performance of machine learning and deep learning. Machine learning methods include methods such as logistic regression, SVM, and random forest. Deep learning approaches include simple deep neural network(DNN), and convolutional neural network(CNN). As a result, random forest recorded the best accuracy among the models. This shows that methods that are known to be more complex, such as DNN or CNN have great performance in unstructured data(image, text), but not always show the best accuracy when comparing with simpler machine learning methods. Noticing that the ensemble based method; random forest outperforms other approaches, appliance of similar tree based ensemble methods may improve the performance. Another classification research [8] held by Suwon city in Korea, was accomplished for the detection of the delinquency of residents in local taxes. To investigate this, random forest, logistic regression, recursive partitioning, and simple neural network models were utilized. In conclusion, the logistic regression model was chosen since it is interpretive. But for sizable data, the usage of efficient and accurate models such as ensemble models will be an improvement.

For tasks regarding regression, a study[13] using the Rossmann Store Sales data[6] from Kaggle, utilized machine learning on time series data. Lasso Regression, Random Forest, Extra Tree, Arima, Stacking of these models were suggested for prediction of the store sales. The stacking method resulted to be most accurate, verifying the excellence of tree based algorithms in regard to time series data. Another regression task [12] used CNN and Recurrent Neural Network(RNN) to predict the stock price for the top 20 issues in KOSPI. RNN had an accuracy for 52 percent. [15] uses an attention method with LSTM to predict the KOSPI stock prices of mobile carriers KT, LGU+ and SKT. Utilizing different types of attention, with element-wise dot product context vectors resulted as the best performance. Both [12] and [15] utilize neural network models to predict stock sales, but with the usage of gradient boosting methods, the results will be human-interpretative showing better reasoning for the data.

As mentioned above, finance data is usually analyzed with simple machine learning methods such as Logistic Regression, Random Forest, and SVM or Neural Network models such as DNN, CNN or RNN. However, the utilization of tree based ensemble methods are rarely executed. Gradient Boosting methods such as XGboost, LightGBM, CatBoost are usually used in areas of environment[9] , housing[10] and science fields[14][5]. These related studies are enumerated in Table 1.

Table 1. Tasks of Past Studies

Task	Goal	Model
Classification	Credit Card Delinquency Prediction[11]	SVM, Ensemble of SVM
Classification	Default Payments of Credit Card Clients in Taiwan[16]	Logistic Regression, SVM, Random forest, DNN, CNN
Classification	Suwon Citizen Delinquency prediction[8]	Logistic Regression, Recursive Partitioning,nRandom forest, DNN
Regression	Rossmann Store Sales Prediction[13]	[Stacking] Lasso Regression, ARIMA, Extra tree, XGBoost, DNN
Regression	KOSPI Stock Price Prediction[12]	DNN, CNN, RNN
Regression	KT, LGU+, SKT Stock Price Prediction [15]	Attention-Bi-LSTM
Classification	PM2.5 in Seoul Prediction[9]	Ensemble of XGBoost
Regression	Seoul Residential Apartment Price Prediction [10]	[Stacking] XGBoost, LightGBM
Regression	Load Forecasting [14]	XGBoost
Regression	Safety Driver Prediction[5]	XGBoost, LightGBM

On contrast of these previous researches of finance data, we implemented tree based machine learning methods to predict future store sales[4]. We will utilize Gradient Boosting based models such as XGBoost, LightGBM, CatBoost which were not preferred in previous finance data analysis. We will also deal with various methods of handling missing data and compare the performance with each model.

2. Statement of goal

Our main objective is to explore methods of handling missing data with the ”Store Sales Data” from Dacon[4] and fit gradient boosting models to make predictions for each individual store. We are trying to answer the following questions: (1) is there a state-of-art method in handling missing data? Also, (2) is there a leading model that generates noticeably accurate predictions? If so, (3) can we predict future three month sales with this method?

3. Methods

The methods section is organized into three sections: Section 3.1 explains the dataset and variables; Section 3.2. presents the preprocess process; and Section 3.3 shows the final features for the dataset.

3.1 Explanations of the data This study utilizes the data from a Korean data competition platform DACON [2], ’Store Credit Card Sales Prediction Competition’[4], which was held from 11, July, 2019 to 21, October, 2019. This data is contributed by FUNDA[3], and the variables are enumerated in Table 2. The period of the dataset is from 1, June, 2016 to 28, February, 2019.

Table 2. Train Set Variables

variable name	Explanation
store_id	Store ID
card_id	Credit Card ID
card_company	De-identified Card Company Name
transacted_date	Date of Transaction
transacted_time	Time of Transaction
installment_term	Installment Term of Transaction
region	Store Region
type_of_business	Type of Business
amount	Transaction Amount

3.2 Preprocess As 31% of the variable ’region’ is missing data, and the variable ’type_of_business contains 60% of missing data, we eliminated these variables. The proportion of missing values for each variable can be seen from the Fig. 3. Additionally, we deleted the variables of the following : ’card_id’,’card_company’,’transacted_time’, due to the irrelevance in the predictions of store sales.

The dataset included zeros and negative numbers in the amount variable. To deal with these irregular amounts, we first deleted the rows including zero as the amount. For the negative amounts, we assumed it as refunding situations, so we matched amounts regarding store_id and card_id and deleted both negative amounts with the identical previous positive amount.

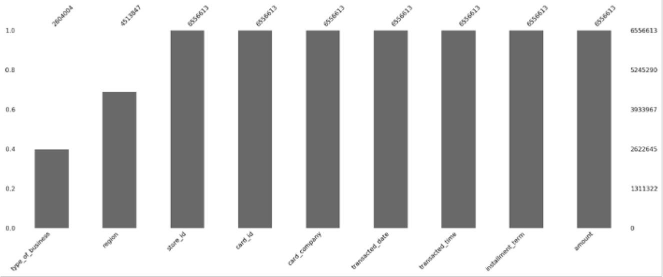

Fig. 1. Missing Data Proportion Plot for the Original Dataset

As the objective is to predict the future three months sales for individual stores, we re-sampled the dataset to a monthly basis. To be specific, we merged the daily amounts into monthly sales for each store.

To emphasize the monthly amounts, we created new variables displaying past sales. We added past three month sales, past four month sales,... past twelve month sales. As this process utilizes the past sales, it produces various missing data. This is where we apply methods for handling missing data, such as: mean imputation, median imputation, linear interpolation and spline interpolation. As the newly created variables are based on a maximum, twelve months past sales, we re-organized the dataset term to June, 2017 from June, 2016.

From the following, Fig 2 we can infer the time difference for each store. Regarding this, we did not consider fixed dates for the future three months. Instead, we considered each store individually, selecting the last three months as the test term. To be specific, store_id 1 has the full term of June, 2017 to February, 2019. So, for this store, we will assume June, 2017 to November, 2018 for our train data, and December, 2018 to February, 2019 for the test data. Another example from Fig 2 is store_id 772, which is expressed as the square lines, and includes data until October, 2018. For this store, the train data will be June, 2016 to July, 2018 and the text data will be August, 2018 to October, 2018. As shown through the Fig 2, each store has variant end dates being impossible to consider the same prediction dates.

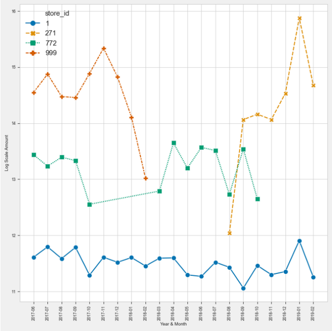

Fig. 2. Difference Between Time Series Plots for Store ID 1, 271, 772 and 999

3.3 Final Variables After the preprocessing, we conclude the following variables as shown in Table 3.

Table 3. Details in Final Variables

Variable name	Explanation
store_id	Store ID
month	Month of Transaction Date
year	Year of Transaction Date
three_month	Total Transaction Amount three months before the transaction date
four_month	Total Transaction Amount four months before the transaction date
five_month	Total Transaction Amount five months before the transaction date
six_month	Total Transaction Amount six months before the transaction date
seven_month	Total Transaction Amount seven months before the transaction date
eight_month	Total Transaction Amount eight months before the transaction date
nine_month	Total Transaction Amount nine months before the transaction date
ten_month	Total Transaction Amount ten months before the transaction date
eleven_month	Total Transaction Amount eleven months before the transaction date
twelve_month	Total Transaction Amount twelve months before the transaction date
amount	Total Transaction Amount on a monthly basis

4. Results

The results section is organized in the following sections. First, explanations of the datasets produced through missing data management and the evaluation metric. Second, detailed information of the models utilized. Third, results from each model: XGBoost, LightGBM, CatBoost. Finally, visualizations of explaining about the predictions made from the selected dataset and models.

4.1 Datasets and Evaluation Metric The datasets we used for the training process is a total of five, which include the dataset with the original missing data, a dataset filled with mean imputation, a dataset filled with median imputation,a dataset filled with linear interpolation, and a dataset filled with spline interpolation.

The metric we used to measure the magnitude of errors is Mean Absolute Error(MAE). It measures the average magnitude of the errors in a set of predictions, without considering their direction. Simply, we calculate the absolute difference of the actual values and predictions and make an average of them throughout the whole dataset. MAE is calculate by the following equation.

$$MAE = \frac{\sum |y_t - \hat{y_t}|}{n}. \tag{1}$$

4.2 Model Utilization With the five datasets in mind, we utilized XGBoost, LightGBM, CatBoost to make predictions for the future sales. To prevent over-fitting issues, we used a stratified five fold cross validation method to maintain every store to be included in the training process. We optimized the parameters for each model by the mixture of random grid search and Bayesian optimization.

Table 4. MAE Test Score Results

	XGBoost	LightGBM	CatBoost
NA	321279.78	320953.13	360355.44
MEAN	323330.41	326836.75	353011.75
MEDIAN	320487.36	325040.74	353409.89
LINEAR	323462.43	323223.69	359235.33
SPLINE	327696.05	323269.50	354742.81

4.3 Model Results The following Table 4 shows the MAE score results for each dataset and model.

For the XGBoost model, there are no internal methods that deal with categorical values, so we composed dummy variables for categorical variables by one-hot-encoding. As we can speculate from Table 4, for XGBoost, the median imputation dataset has the best accuracy, followed by NA dataset. For LightGBM, the combination of the NA dataset has the acute accuracy, which is followed by spline interpolation dataset. Lastly, for the Catboost model, the mean imputation dataset is the finest match.

According to Table 4, we can conclude that the combination of XGBoost and the median imputation dataset has the first-rate MAE score. This is trailed by solution of the NA dataset and LightGBM model. The Catboost model has the lowest accuracy regarding all three models. But for comparison, we selected the combination of the mean imputation dataset and Catboost model. By the following Fig 3, we can easily compare the error rate of each models.

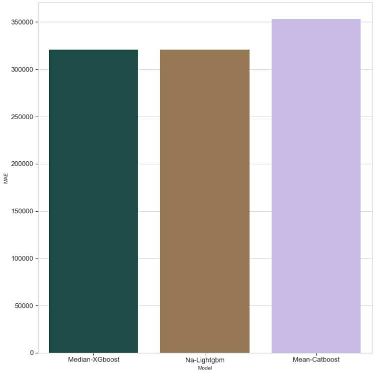

Fig. 3. MAE Barplot for the Final Models

5. Discussion

Is there a best combination regarding the methods of handling missing data and models for training the data?

From Table 4, we can conclude that the XGBoost model with the median imputation dataset has the lowest error rate. When comparing the Median-XGboost with the

NA-LightGBM combination, we can see a very limited difference between each model's MAE. This results indicates that there is no absolute model and missing data management for the dataset. Also, from the following Table 5, we can infer that XGBoost does conclude the highest accuracy, but the time measurements show a different aspect of the model. Compared with the LightGBM model which only depletes 9 seconds to make predictions, XGBoost consumes severe amounts of time, over 3000 seconds. With the small-scale difference in MAE, and the large-scale difference in time consumption, we can conclude that the final missing data method-model combination is not implicitly detectable.

Table 5. Comparisons of MAE with Time Measurements

Dataset	Model	Mae	Time(s)
Median	XGBoost	320487.36	3537.50
NA	LightGBM	320953.13	9.70
Mean	CatBoost	353011.75	35.70

Can we predict the future store sales from this combination? With the final model selections from Table 4, we are capable of visualizing the predictions of thirty random selected stores. The following Fig 4 shows the visualization of future three months store sales. The actual amount is expressed with x-lines, the predictions made by the XGBoost model with median imputation is the triangle-lines, the predictions made by the lightGBM model with NA dataset is shown with the rectangle-lines, and the predictions made by the Catboost model with the mean imputation dataset is visualized with the diamond-lines. From the graph visualization, we can easily comprehend that the XGBoost and LightGBM model gradually follows the actual sales.

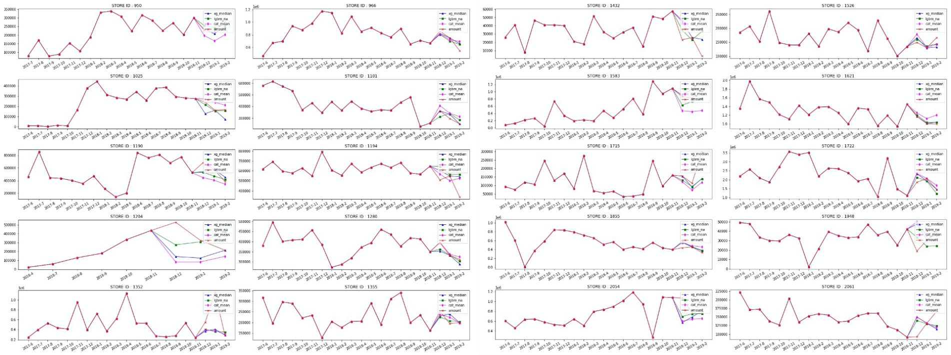

Fig. 4. Comparisons of MAE through Graph Visualization

6. Conclusion

This paper focuses on the methods of handling missing data and the accuracy associated with the utilization of gradient boosting models. We were able to compare the accuracy for each imputation or interpolation method when combined with machine learning models. With the missing data in the store sales data, we created five datasets: NA(dataset with the missing data), mean imputation, median imputation, linear interpolation, and spline interpolation. These datasets were utilized with three gradient boosting models: XGBoost, LightGBM and CatBoost. We conclude that the combination of XGBoost and mean impuatation is state-of-art but regarding time efficiency, this combination is incomparable with the LightGBM model with the NA dataset, as it is 400 times slower regarding the latter.

Nowadays we can easily find applications of data in fintech fields. Such as: convenient data driven purchase mechanisms, P2P loans, and financial item recommendation.[1] With these services in mind, we believe that this paper can be a background for fintech enterprises to display new methods in evaluating and aiding future customers. With the help of machine learning, fintech firms can enforce methods of personalized service for each customer. Through the individual aspects of the model, these firms can easily evaluate customers in a more accurate method than the monotonous past approaches which only used the customer's basic information.

7. Limitations

The variable tally of the original dataset was restricted to only daily sales. As there were few features to use, less creative approaches of machine learning methods were available to consider in the study. With more variables to examine, a more accurate and inventive model may be built.

8. Future work

Perhaps considering more complicated methods such as: Deep Neural Networks, Long short term Memory(LSTM) methods, or time series approaches such as SARIMA or ARIMA, may improve the accuracy of the result. Future analysis not only with deep learning and time series methods but also with simple stacking approaches may show improvements in the error rate.

References

1. Banksalad magazine, [Online]. Available: https://banksalad.com/contents/C
2. Dacon.korea data competition platform., [Online]. Available: Available:https://dacon.io/
3. Funda-korea p2p finance finance enterprise, [Online]. Available:https://www.funda.kr/v2/
4. Dacon: Card sales prediction contest, [Online]. Available: https://dacon.io/competitions/official/140472/overview/.
5. Jang, S.I., Kwak, K.C.: Comparison of safety driver prediction performance with xgboost and lightgbm. In: Proceeding of Korea Institute of Infomation Technology Conference. pp. 360–362. Seoul, Korea (2019)

6. Kaggle: Rossmann store sales, [Online]. Available: https://www.kaggle.com/c/rossmann-store-sales
7. Kaggle: Uci credit card dataset, [Online]. Available: https://www.kaggle.com/uciml/default-of-credit-ca rd-clients-dataset
8. Kim, E.C., Yoo., B.J.: Case study of analyzing credit information and public big data : Development of prediction model for probability of collecting tax in arrears and enhancement of tax information system. In: Proceeding of 58th the Korea Society of Management information Systems Conference. pp. 6–12. Seoul, Korea (2018)
9. Kim, H.: The prediction of pm2.5 in seoul through xgboost ensemble. Journal of the Korean Data Analysis Society 22(4), 1661–1671 (2020)
10. Kim, I.H., Lee, K.S.: Tree based ensemble model for developing and evaluating automated valuation models : The case of seoul residential apartment. Journal of the Korean Data Information Science Society 31(2), 375–389 (2020)
11. Kim, J.W., Jhee, W.C.: Credit card delinquency prediction model based on data mining approach. In: Proceedings of the Korea Intelligent Information System Society Conference. pp. 232–239. Korea Intelligent Information Systems Society, Seoul, Korea (2011)
12. Lee., J.: Stock price prediction model using deep learning. Masters dissertation,Soongsil University, Soongsil University, Seoul (2016)
13. Pavlyshenko, B.M.: Machine-learning models for sales time series forecasting. Data 2019 4(1), 15 (2019)
14. Y. G. Lee, J.Y.O., Kim, G.B.: Interpretation of load forecasting using explainable artificial intelligence techniques. The Transactions of the Korean Institute of Electrical Engineers 69(3), 480–485 (2020)
15. Y.J.Kim: Attention Mechanism Based Using Bi-Directional LSTM Stock Price Forecasting Model. Masters dissertation,Hanbat National University, Hanbat National University,Daejeon (2019)
16. Yoon, J.M.: Effectiveness analysis of credit card default risk with deep learning neural network. Journal of Money Finance 33(1), 151–183 (2019)