**Predicting Future sales**

**Advanced machine learning Assignment**

SSID:2021497

MOD006566

**Table of contents**

1. Introduction

2. Literature Survey

3. Data Analysis and Data Pre-Processing.

4. Models

4.1 Linear Regression

4.2 Polynomial Regression

4.2.1 Lasso regression

4.2.2 Ridge regression

4.3 Gradient Boosting Regression

4.4 Extreme Gradient Boosting

4.5 LightGBM

4.6 Linear Ensemble models

4.7 other models

5. Evaluation of model

6. Kaggle Submission

7. Legal, Ethical and Privacy concerns

8.Conclusion and future improvements

9.References

**1. Introduction**

The main objective of this work is to detect the trend in the sales so that it can help for the organisation to improve their business. In today’s world data plays a major role in getting the profits to their organisation particularly in e-commerce domain. This is where machine learning techniques comes into picture for forecasting sales so that it can add value to the company.

Predicting future sales is a Kaggle challenge and the data is based on daily historical sales data. And the data is provided by the largest Russian software firms – 1C Company. The use case here is of regression type.

Data mining is also one of the technique to get the insights or hidden patterns from the large volumes of data. Many organisations faces several problems to select the accurate machine learning algorithm or data mining algorithms to predict the future sales.

In our project I have applied different machine learning algorithms for the future sales prediction and selected the best algorithm by comparing with the losses of the other machine learning models.

**2. Literature Survey**

S. Cheriyan, S. Ibrahim (2018) used three models named generalized linear model, Decision Tree Regressor and Gradient boost Regressor. In this paper S. Cheriyan, S. Ibrahim experimented more about the data analysis part like performed exploratory data analysis to get the insights from the data and also explore the trends in the sales. They also performed outlier detection which is the key for impacting the performance. Out of Generalized linear model, Decision Tree Regressor and Gradient boost Regressor, Gradient boost Regressor performs better when compared to remaining algorithms and the accuracies are 64%, 71% and 98% respectively. So, the authors of this papers uses gradient boosting Regressor as final model because of better accuracy and minimum error rate.

Aneesh tony , Pradeep Kumar (2021) in their paper describes the problems faced in retail industry to gain the profits. The main problem where every business man focused in the retail industry is to improve the sales of the product. Mostly it will depends on the some of the factors like customer feedback about the product items sold etc. For checking these things every time is bit more difficult task to avoid these problems we used machine learning techniques. In this paper author used Linear Regression, Decision Trees and Random Forest algorithms for predicting the sales and the authors used root mean squared error for evaluating the models. Random Forest Regressor performs better when compared to linear regression and decision trees with less root mean squared error and less mean absolute error.

From the above two papers it is understood that Ensembling models like random forest and gradient boosting classifier performs better.

Bohdan M. Pavlyshenko (2018) published his paper on sales time series forecasting using machine learning models. In this paper Bohdan M. Pavlyshenko proposed the novel approach for forecasting the sales in the future. We basically use single machine learning model for the forecasting purpose. Here in this paper authors used stacking approach which is nothing but Ensembling different machine learning algorithms and build into a single algorithm so that it is robust to anything and the results obtained are accurate.

Pavlyshenko, Bohdan M (2019) utilised the many single models, where most of the results produced were based on XGBoost machine-learning algorithm that could achieved the sought after understanding for the predictive sales model. For the second level of the yardstick, they employed two models from Python scikit-learn package—ExtraTree model and linear model in addition to Neural Network model for the best predictive sales modelling that is the need of this paper. The results of the experimentation from the second level yardstick were stated with the required and summed up weights by basing the same on the third level yardstick. The construction based on the new features were achieved that was the foremost feature of the modelling. Basing it on aggregating target variable and its lags that were grouped with different and variable factors. In terms of evaluation, the utilisation of Relative Mean Absolute Error (MAE) was done making it the absolute necessity of the modelling technique and the measures that can be employed.

**3. Data Analysis and Data Pre-processing:**

There are 6 files in our dataset. They are Item\_categories.csv, items.csv, sales\_train.csv, sample\_submission.csv, shops.csv and test.csv. Out of all these files sales\_train.csv is used to train the model and the remaining data files are used to get the insights.

The data in the sales\_train.csv are shown in the below snap.

Table

Description automatically generated

The below snapshot provides us the information about the number of records and attributes present in those 6 files.

Text, letter

Description automatically generated

From the above figure we can say that in training data the number of records are 2935849 records and the attributes or features or columns are 6 columns. Regarding test dataset the number of records are 214200 records and 3 columns.

The Statistical analysis of the train data can be obtained by using .describe (). The results for the statistical analysis can be shown in the screenshot.

Text, table

Description automatically generated

Graphical user interface

Description automatically generated

The price items range is mostly in between 200 to 400. There is also a rare chance of high price items to be sold as per the analysis from the above graph. The number of unique items present in the stores are 21807, Number of Unique categories for Items Present in the stores available are 84 and the number of unique shops are 60

Text

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generatedThe busiest days for the shops is initial days of the month, the busiest month for the shop is December and the busiest year for the shops according to our data is 2013. These conclusion are drawn from the below figures.

Chart, bar chart

Description automatically generatedGraphical user interface, website

Description automatically generated

Fig . Heat map

It can be concluded from the above figure that Year and date\_block\_num which consists of number for each month are highly correlated. Some of the correlations can be analysed from Heat map above.

Next step is to check for the outliers of ‘item price’ and ‘item count day’ attributes. For that I used plotting for both the attributes and from the graphs and concluded that the item price is greater than 10000 is as an outlier and for item count per day if the items sold per day is greater than 1200 is also an outlier. To select the records without outliers and also one more thing to notice is some of the prices in the data are negative values; I have replaced those negative values with median.

In Data Pre-processing, checking numbers in training and testing are equal or not is important. It can lead to underfit or overfit of model if the values are not distributed well. After checking it is found that 363 items are missing. As it is not huge number I have to ignore and continue with other data-pre-processing techniques.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

In the shops dataset the ‘shop name’ column is the combination of city and the kind of shop. I have to remove the unwanted symbols (‘!’) from the shop name and separated city from the shop name attribute and also assigning code to the city. And code for city label is added to dataset.

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Similarly, item categories data set the ‘item category name’ attribute having the combination of type of the category and sub types. I have extracted the data from ‘item category name’ attribute and create some attributes like ‘cat type’, ‘cat type code’, ‘sub category type’ and ‘sub category type code’ to store the extracted data. Below is the code for the above process.

A picture containing scatter chart

Description automatically generated

Final train dataset is join of sales, shops, items, item\_categories dataframes.

Text

Description automatically generated

**4. Models:**

In this project, I have applied basic level machine learning algorithms like linear regression and regularization methods like ridge regression and lasso regression to advanced regression machine learning algorithms like gradient boosting regressor, Decision Tree Regressor, Random Forest Regressor, Extreme gradient boosting Regressor, light gradient boosting Regressor.

Before feeding the data to the model we need to separate the data into training data and validation data. Since the output for sales prediction starts from Nov, 2015 i.e. date\_block\_num = 34 so I have added that to our test to make sale prediction.

Text

Description automatically generated

**4.1 Linear Regression**

After segregating the data firstly, I applied linear regression model to the training data since it is a basic model there is no hyper parameters for linear regression algorithm. The code for the linear regression algorithm is shown in the below screenshot.

Graphical user interface, text, email

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Linear regression is a basic algorithm, and it is sensitive to outliers and also if the assumptions of the linear regression doesn’t meet in that case also it won’t perform well. So, to make the model robust I have applied regularization methods of linear regression called ridge regression and lasso regression.

**4.2 Polynomial Regression**

**4.2.1 Lasso Regression:**

Lasso regression is one of the regularization technique. Lasso stands for “Least Absolute Shrinkage and selection operator”. It is also called as L1 regularization the main objective of Lasso Regression is to reduce the Overfitting scenarios. In Lasso Regression there is one penalty to the magnitude of the co-efficient to the cost function (‘ ƛ’).

The main advantage of Lasso regression is it is not only used to reduce the over fitting cases but also used in feature selection. Below is the code for the lasso regression without applying any hyper parameter tuning.

Graphical user interface, text, application

Description automatically generated

I have used grid search cv to select the best hyper parameters for selecting the best alpha value and also the best score in terms of mean absolute error. I have used k fold cross validation technique along with ‘grid search cv’. The below code depicts the hyper parameter tuning using grid search CV.

Graphical user interface, text, application, email

Description automatically generated

**4.2.2 Ridge Regression:**

Ridge Regression is also called L2 regularization. It is also used to reduce the Over-fitting by changing the cost function with the little modification of adding penalty equivalent to the square of the magnitude of coefficients. By adding this we can minimise the cost function.

I have implemented ridge regression by taking several alpha values and performed hyper parameter tuning with the help of grid search CV along with that we apply k fold cross validation. The code below shows the best parameters in this case the best alpha value and the best score in terms of mean absolute error.

Graphical user interface, text, application, email

Description automatically generated

**4.3 Gradient Boosting Regressor**

Gradient Boosting Regressor uses gradient descent to reduce the loss. So I have implemented using with RandomizedsearchCV with 5 fold validation and parameters include max depth of 3,4 and 5, and number of estimators as 50, 100, 200 with 0.01,0.5, 0.1 learning rate.

A picture containing graphical user interface

Description automatically generated

I was able to get better performance by changing the number of estimators and train the model again to see the difference as shown in the below figure. I have calculated mean square error to see that difference and submitted which is working better than others.

Graphical user interface, text, application, email

Description automatically generated

**4.4 Extreme Gradient Boosting (XGBoost):**

Extreme gradient boosting is an Ensembling technique which falls under boosting type. The base estimators or models for the extreme gradient boosting technique is Decision Trees. The main advantage is it minimizes its loss gradient whenever the model undergoes training. XGBoost possess lots of parameters to tune the model. XGBoost itself can prevent the Overfitting scenarios and also it is robust to the outliers. XGBoost is used mainly to reduce the bias not only XGBoost all boosting algorithms aims to decrease the bias.

I have implemented XGBoost algorithm by taking some of the hyper parameters by taking maximum depth as 8, number of estimators or decision trees as 1000, and by taking some other parameters. Below screen shot shows the code for the XGBoost model training.

Text

Description automatically generated

The score which we are getting after submitting our submission file to Kaggle is 1.1244 which is better than previously used algorithms.

**4.5 Light Gradient Boosting Machine (LGBM):**

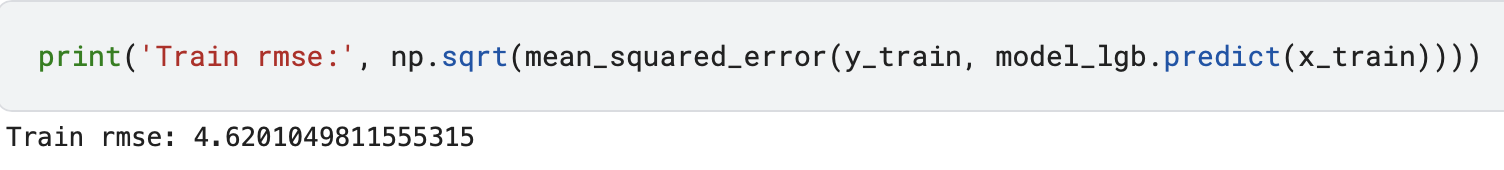
Light gradient boosting machine is fast, distributed and high-performance gradient boosting framework which uses decision tree algorithm as the base estimator. The basic difference of light gradient boosting machine over XGB is 7 times faster than the XGBoost algorithm regarding the performance there is slight difference in terms of accuracies.

As similar to XGBoost algorithm, LGBM is also having more hyper parameters to tune the model.

I have implemented Light Gradient Boosting algorithm by taking some of the hyper parameters by taking maximum depth as 8, number of estimators or decision trees as 200, learning rate as 0.03, number of leaves as 32 and by taking some other parameters. Below screen shot shows the code for the Light gradient boosting machine model training with some parameters.

Text

Description automatically generated



**4.6 Linear Ensemble of models**

I have created a second layer where I have used submission files of LGBM and Gradient Boosting, Extreme Gradient Boosting, and Random forest Regressors. Since the performance of these models is high when compared to others. Second layer of stacking include simple linear regression where I have used part of one model with part of other which gave log score of 1.04265 which is less than LGBM regressor log score on test data but more than single Gradient Boosting regressor model. Like this, I have made two more combinations; the scores for each combination is listed in the table below.

|  |  |
| --- | --- |
| **Ensemble Model** | **Kaggle submission score** |
| LGBM+GradientBoosting | 1.04265 |
| LGBM+XGB | 1.04747 |
| LGBM+Randomforest | 1.05364 |

**4.7 Other models**

Decision Tree regressor and Random Forest Regressors are the other two regressors. Below are the figures showing cross validation scores for both regressors. Random forest regressor performance is better than Decision tree on test data. I have used cross validation of 10 fold in Decision Tree and Random Forest.

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Fig showing validation score of Decision Tree and Random Forest

In the above figure, model3, model4 are trained on Decision tree and Random Forest. This figure contrasts the validation score of both the models.

**5. Evaluation of model**

Error metrics are important to evaluate model, how well the model is trained and helped in predicting outputs. For Evaluating Metric each model I have used Root mean square error and listed the error values in the below table.

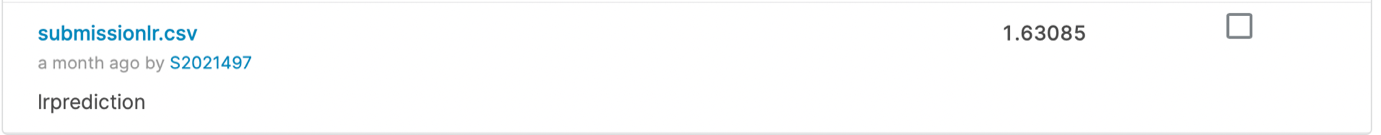
|  |  |  |
| --- | --- | --- |
| **Model** | **Kaggle submission score** | **RMSE Scores** |
| **Linear Regression** | 1.63085 | 3.86710 |
| **Lasso Regression** | 1.54595 | 3.88978 |
| **Decision Tree** | 1.21805 | 3.11413 |
| **Random Forest** | 1.14635 | 3.11713 |
| **XGBRegressor** | 1.12444 | 3.10678 |
| **Gradient boosting** | 1.07295 | 2.89765 |
| **LGBMRegressor** | 1.03349 | 2.11513 |

**6. Kaggle submission**

For Kaggle submission i have submitted the fourteen-submission file created by 7 machine algorithms and two combination models. The machine learning algorithms used are Linear Regression, Lasso Regression, XGB Regressor and LGBM Regressor algorithms. Out of all Machine learning regression algorithms Light Gradient boosting Regressor algorithm performs well with less mean squared error as 1.03349 when compared with the remaining algorithms 1.63085 for Linear Regression, 1.12444 for extreme gradient boosting Regressor, and 1.07295 for Gradient Boosting Regressor. With combination of two models I have achieved 1.04265 with LGBM and Gradient boosting regressor which was the second highest after LGBM. All the remaining scores can be compared from the below figure.

Graphical user interface, text

Description automatically generatedGraphical user interface, application

Description automatically generated

**7. Legal, Ethical and Privacy concerns**

Firstly we will think from the side of retailer, the one who is a retailer must protect his/her data confidentially in order to get rid of data breach. So if he want to works with his own data to enhance his own business, it might be effective but he should protect his customer’s data. If the data of the customers get breached, it might be a legal issue. While collecting data from the customers in the form of feedback or any else. The questionnaire must be in the form of ethical i.e., by simply ignoring the personal or unwanted data. The retailer must focus on privacy of the citizen while preparing or collecting data. Means simply protecting user’s sensitive data.

**8. Conclusion and future improvements**

In this project we implemented powerful machine learning techniques like extreme gradient boosting and light gradient boosting in order to predict the future sales. Before applying machine learning models a more detailed data analysis is performed on the future sales dataset. As per the results obtained from the above section light gradient boosting machine Regressor having low error or better score compared to remaining regression algorithms.

For further improvements, we can gain some more techniques by following the Kaggle winners and add some advanced machine learning algorithms by applying hyper parameter tuning. This may decrease our mean squared error.

**References**

* Aneesh tony., Kumar, P., & Rohith Jefferson, S. 2021. *A Study of Demand and Sales Forecasting Model using Machine Learning Algorithm*. Psychology and Education Journal, 58(2), 10182-10194.
* Beheshti-Kashi, Samaneh & Karimi, Hamid & Thoben, Klaus-Dieter & Lütjen, Michael & Teucke, Michael. 2015. *A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering*: An Open Access Journal. 3. 154-161. 10.1080/21642583.2014.999389.
* Huang, W., Zhang, Q., Xu, W., Fu, H., Wang, M., & Liang, X. 2015. *A Novel Trigger Model for Sales Prediction with Data Mining Techniques.* Data Science Journal, 14.
* S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, 2018. *Intelligent Sales Prediction Using Machine Learning Techniques*, 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115.
* Swami, Devendra & Shah, Alay & Ray, Subhrajeet. 2020. *Predicting Future Sales of Retail Products using Machine Learning.*
* B. Pavlyshenko, 2018,*Using Stacking Approaches for Machine Learning Models*,  *IEEE Second International Conference on Data Stream Mining & Processing (DSMP) IEEE*, pp. 255-258.
* Pavlyshenko, Bohdan M. 2019. *Machine-Learning Models for Sales Time Series Forecasting.* [e-journal] Data 4, no. 1: 15. <https://doi.org/10.3390/data4010015>