ML Final Project Presentation

Group 2 – Revenue Forcasting For Businesses

Utkarsh Rajauria 2019214
Amit Chaurasiya 2019142
Aadarsh 2019131

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Problem Statement and Motivation

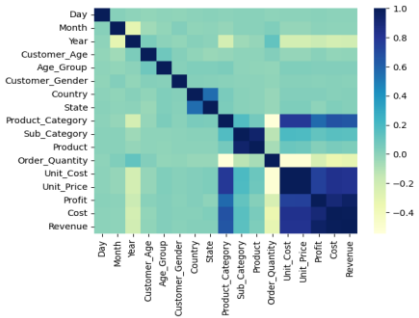
1. Business face Lots of losses due to the unpredictable nature of sales in the future. Therefore there is a need to have a rough estimate of sales in the future so that they can have adequate stock of items.

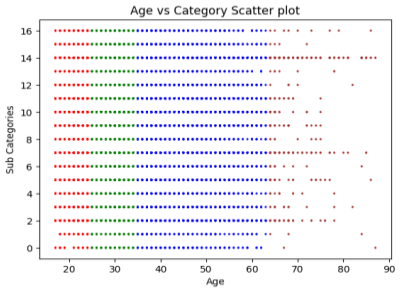
2. So, We will try to solve the problem with the help of state-of-the-art Machine Learning Techniques.

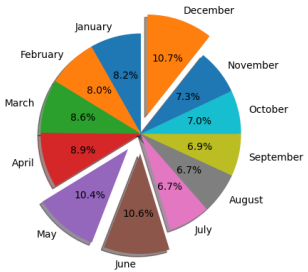
Exploratory Data Analysis and Data Processing

Exploratory Data Analysis, Preprocessing, Normalising and Modelling the data

- Dropped null values and duplicate rows.
- Converted date and time attributes from string to date time type and created it as an index column.
- Used Correlation matrix to find the redundant features which helped us to reduce features from 15 to 10.
- Used PCA to reduce overfitting.
- Used MinMaxScaler( ) for normalization and converting data values within the range (0,1).
- RMSE has been used as the Evaluation Metric.





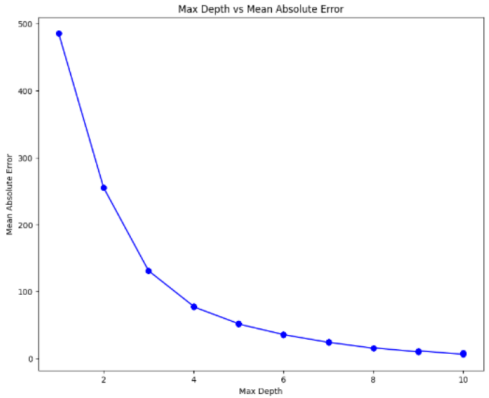


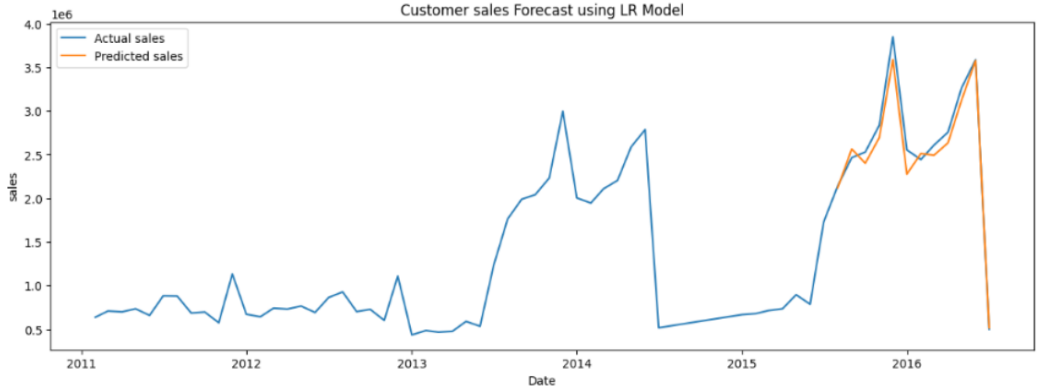
Exploratory Data Analysis and Data Processing

- **Analysis after performing the EDA:**
- There were no null values found.
- After finding the correlation matrix, we removed the highly correlated features.
- Gender is not an important feature as there was no significant difference found in sales.
- Bike category has the highest revenue.
- The highest sales were noticed in December, June, and May.
- Most of the subcategory’s sales drop for customers aging > 65.

Model Selection, training and Interpretation

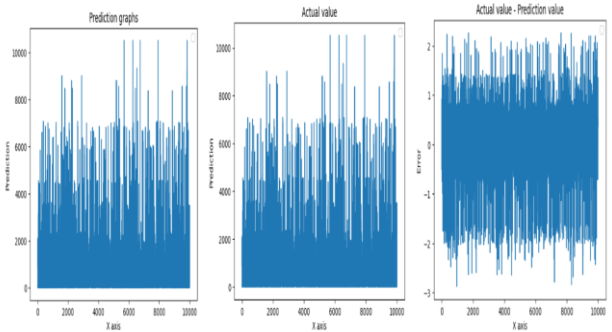
- **Linear Regression:** The data was splitted into training(70%) and testing(30%). After removing the redundant features from our dataset, 10 of them were left. Thereafter we applied the linear regression model. Here are the observations:
  RMSE: 56.0017613663
- **Decision Tree:** We have used the decision tree regressor here to explore for better results compared to the previous linear regression model. Here we are first dividing our dataset into train(60%), validation(20%) and test(20%) sets.
  RMSE: 17.552048499545403
- **Random Forest with K fold cross-validation:** We are trying to find a better RMSE but after applying K fold cross-validation our RMSE gets increased, where the value of k is 5.
  RMSE: 44.31867454822607





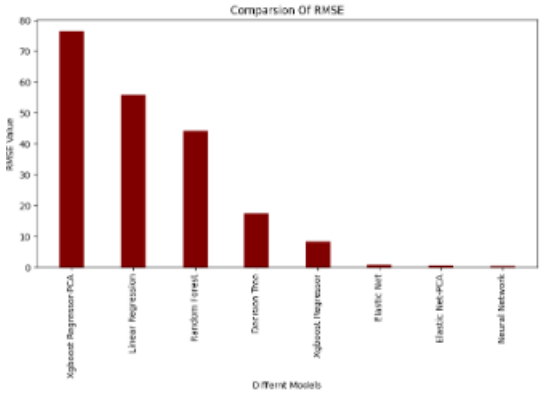
Model Selection, training and Interpretation(Contd.)

- **XGBoost Regressor with PCA:**
  We have used **Grid Search** using the **GridSearchCV** module to find the best hyper parameters. Based on the outcome, the XGBoost regressor without PCA has performed significantly better than the XGBoost regressor with PCA. One of the reasons is that the PCA transformation is not preserving enough information from the original features, leading to a loss of predictive power.
  RMSE : 76.49888789902742
- **Elastic Net with PCA:**
  As we saw earlier, the LR model was over fitting our data which led to its bad performance. Hence, to reduce the over fitting issue we have used Elastic Net regularization technique which is a combination of both L1 and L2 regularization. As showcased in the results of every model, Elastic Net with PCA has given the best results for our dataset.
  RMSE : 0.6139800567406719
- **Neural Networks:**
  RMSE : 0.3377558452724246
  In the neural network model there is one hidden layer with 32 neurons and a ReLU activation function, and an output layer with one neuron and a linear activation function. The model is trained for 100 epochs with a batch size of 32 using the fit method of the model.



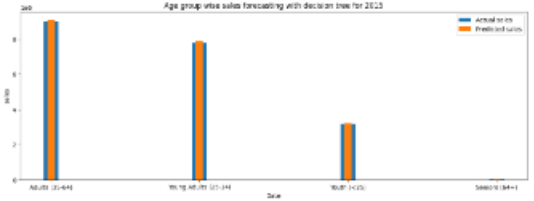
Inferences drawn from results and Conclusion

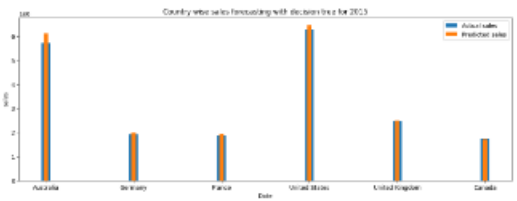
- In our analysis we found out that linear regression was not able to perform really well with over fitting being one of the causes. Random forest is able to perform a little better with cross validation applied. Decision tree further improved the performance with the best hyper parameters found with the Grid Search. Further we also tested the XGBoost Regressor model for trying gradient boosting and found its best hyper parameters through grid search. Another drastic improvement was found with applying ElasticNet with PCA where dimensionality was reduced to 5 which reduced the complexity of the model and combination of L1 and L2 regularization techniques were applied to further prevent over fitting and to get a much better fit on test data. Finally the best performance was found using Neural Networks.
-



Novelty, Real-World Application, and Future Scope

- We have applied techniques like Grid search, gradient boosting and Principal Component Analysis to optimize our results which are not being worked upon much in the domain of sales forecasting. We have also added the category wise predictions below to show the versatility of our trained model.





Code And Report Link

REPORT LINK :
https://drive.google.com/file/d/1BLMIMRwTgyQwHaDeGjeOkRexGWnFJ5Cv/view?usp=share_link

CODE LINK :

https://colab.research.google.com/drive/1wQigv5ZZLc0JbE48cB1j2YOSVJ-We_Zh?usp=sharing