Project Meeting Minutes - Week 1 – 12

Project Name: Multi Mart Total purchases & Loyalty Card forecasting

Below is the summary of the analysis/visualization project we completed in 3rd semester.

The project focused on visualizing and analyzing sales and revenue data of Multi Mart Retail store spanning the period of 2019 to 2023. The primary objective was to provide a comprehensive understanding of sales performance, revenue generation, and customer behavior analysis to support informed decision-making, particularly regarding the potential expansion of its Loyalty Card program into new regions.

Git Hub Project Repository: DAB_Grp7_Capstone_Project

Week 1 & 2:

Tasks:

- Researched on LSTM and Time Series concepts in Machine Learning
- Researched on existing dataset to check whether it can be used for machine learning use case.
- Discussed with the Professor to provide another data set for sales forecasting use case.

Follow up Discussion:

- Features which can be used from the Existing data set to implement Machine Learning models for forecasting/predicting below potential metrices:
  - Revenue
  - Frequency and Revenue
  - Churn
- Need to analyze the correlation between multiple features to identify the ones most suitable for Machine learning use case.
- Possible models to focus on: Regression, Classification

Week 3 & 4:

Tasks:

- Analyzed all the features in the data set to identify all use cases related to revenue, frequency & totalpurchases prediction.
- Converted below text columns to numerical using label encoding.
  - referralsource
  - responsetolastcampaign
  - preferredpaymentmethod
- Worked on data transformation and created below new columns from existing columns to add additional features to the data set
  - avgpurchasevalue
  - tenure
  - Recency
  - avgtimebetweenpurchases
- Identified new correlation using the above newly created and converted columns.
- Above columns were discarded later on as there were no fruitful results or correlations of these columns with other existing features.

Follow up Discussion:

- A good correlation of 48% was observed between Totalpurchases – frequency, so these two can be used as part of model to identify total purchases made by a customer.
- Variables need to be identified to look for predictions of total purchases.
- one Hot encoding of columns is suggested instead of label encoding for below columns:
    - referralsource
    - responsetolastcampaign
    - preferredpaymentmethod
- Feature importance analysis to identify the features which are important as part of this prediction.

<u>Week 5, 6 & 7:</u>

Tasks:

- Converted below text columns to numerical using one hot encoding:
    - referralsource
    - responsetolastcampaign
    - preferredpaymentmethod
- Identified new correlation using the above converted columns.
- Feature importance analysis to identify the features which are important for predicting target variable.
- Created new attribute purchases_category by categorizing the totalpurchases into three categories High/Medium/low and identified correlation between the new attribute with the existing columns.
- These attributes were only created to try out a potential idea of segregating the customers on the basis of the number of purchases that they make and were later discarded as there was no such significance with or without the column.

Follow up Discussion:

- Compare model performance using all the features in dataset vs top 4 or 5 features using R2, MSE, MAE and RMSE metrics.
- Perform Principal component Analysis (PCA) on the model as it will give a new dimension that will capture the variations in target and thus can be used for selecting the best features.
- Compare outputs of PCA with feature importance as this comparison will help select the features.
- If the outcomes for above comparison are different then build different models for both
- Run 4-5 models using PCA and feature selection.
- Use models like Gradient Boost.
- Use lazypredict to compare all possible models at once.

<u>Week 8 & 9:</u>

Tasks:

- Used lazy predict to compare different models' performance on dataset using total purchase as target variable.
- For total purchase, top two performing models were GradientBoostingRegressor and RandomForest with R-Squared value of 0.34 and 0.32. Other models were underperforming compared to these two.
- Identified new target column "response_to_last_campaign" as per the correlation matrix after one hot encoding.
- Used lazy predict to compare different models' performance on dataset using "response_to_last_campaign" as target variable.

- For response_to_last_campaign, most of the models were overfitting with R-Squared value of 1. Hence it was discarded.

Follow up Discussion:

- Run lazy predict using only top 5-6 features for total purchase as target variable.
- Use hyperparameter tuning and feature importance on the new model for improvement in model performance.
- Run lazy model using only top 5-6 features for response_to_last_campaign as target variable.
- Analyze the correlationship between variables after one-hot encoding and check which other variables can be considered as target variable.

<u>Week 10 & 11:</u>

Tasks:

- Interim Presentation
    - Prepared presentation to showcase activities completed till 9th week in the project and upcoming activities.
    - Showcased a highlight of the final product.
- As per the output from Lazy Predict, tried GradientBoostingRegressor and RandomForestRegressor using totalpurchases as a target.
- Further performed GridSearch to identify the combination of best possible hyperparameter values to tune the model.
- Researched on the correlation of variables again to identify if there are any other variables which can be used for prediction.
- Identified a high negative correlation between hasloyaltycard and feedback score of -0.87 and applied Machine learning models for predicting hasloyaltycard.
- This will be helpful to predict whether the customer will buy Loyalty card or not. So the store management can decide on the basis of certain parameters to suggest a loyalty card to the customer.
- Feature importance performed to be used with hasloyaltycard and below features were used to predict hasloyaltycard.
    - Totalrevenue
    - Totalpurchase
    - Frequency
    - Feedbackscore
    - Discountused
    - Referralsource
    - Churnindicator
- The following metrics were calculated to examine the model performance.
    - R-squared value using Random Forest model: 0.7845644755437593
    - Precision: 0.9067164179104478
    - Recall: 0.9067164179104478
    - F1 Score: 0.906716417910447
- Prepared the design for final product, User Interface web portal.
- Identified the columns or questions which need to be added in the UI.
- Researched on Machine learning model deployment process using joblib and connecting the model data to user interface using flask.

Follow up Discussion:

- Since the target variable hasloyaltycard has a negative correlation with feedbackscore, which means the customer with no loyalty card are giving good feedback whereas the ones holding are giving low feedback scores.
- And it was observed that the data has more customers without loyaltycard (~8632) then with loyaltycard (~1368)
- Data is imbalance as per the above stats which is influencing the accuracy score.
- Need to identify statistical techniques such as k-fold Cross-Validation to balance the data and then apply Machine learning models.
- Also, another method was discussed to reduce the data set to 4000 records and apply Machine learning models on it to get the accurate scores.

Week 12:

Tasks:

- Applied Random Under Sampling Method and k-fold Cross Validation method of handling imbalanced data after research.
- No major differences in r^2 values were observed even after applying the two methods for our dataset in our model.
- Hence, Loyalty Card prediction could also be used as one of the use cases for our dataset.
- Finalized target variable, model to be used and the features to be used to make these predictions.
- Deployed Machine Learning model for the prediction of Loyalty Card using ‘joblib’.

Up-Coming Tasks:

- Predict total purchases using Label Encoding technique instead of one-hot encoding to reduce additional data preprocessing efforts.
- Final deployment of the model for total purchases prediction and building WEB UI to display the prediction results to the business.
- Building WEB UI for Loyalty Card prediction to help business predict whether to suggest a Loyalty Card to a customer.
- Final Presentation and report to present the overall project.