Data Mining and Machine Learning Assignment 1

Ankan Kar (MCS202303), Suneet Patil(MCS202315)

March 2024

1 Introduction

As part of the first assignment of Data Mining and Machine Learning course, we worked on two tasks. First was prediction of churn in customer dataset and second was predicting gender and rating of a customer based on supermarket sales dataset. In this report we aim to describe the methods deployed by us for understanding both the datasets, detecting presence of outliers in datasets and training models on the datasets. We decided to use Python language because of our familiarity with libraries used for performing all the operations.

2 Customer Churn

2.1 Exploratory Data Analysis

The Customer Churn dataset has 1000 entries, for each record the dataset has 15 features. For each customer, the dataset contained information about Age, Gender, Annual Income, Total Amount Spend, Years since the individual was a customer, Number of purchases made, Number of returns made, Average transaction amount, Number of support queries, Satisfaction score, Days since the customer made their last purchase and whether they have opted for emails and promotions. We understood Correlation of all features present in the dataset by using Heatmaps which were generated using the seaborn library. Then we performed one hot encoding on the categorical columns *Gender*, *EmailOptIn* and *PromotionResponse* in the dataset, which helped in the machine learning model for understanding and extracting data present in these features. Then we scaled the dataset using standard-scaler, this ensured that each feature had standard deviation of 1 and mean of 0.

2.2 Outlier Detection and Removal

For checking presence of outliers, we visualized the data using box-plot, dis-plot and violon-plot present in seaborn library. Then, the dataset was scanned using DBScan to check for presence of any outliers. This ensured that the dataset did not contain any outlier data.

2.3 Train Test Splits

The dataset was split in ratio of 80:20, 80 percent of dataset was used for training and 20 percent was used for testing.

2.4 Models Implemented

Initially, we trained a Adaboost Classifier using decision tree as base model. The best classifier was selected on basis of it's training score which was calculated for a range of models on basis of hyper-parameters. Score of the classifier was 61.3 percent on training data and 60 percent on test data

The we attempted to train a Random Forest Classifier on normalized data with same train test split of 80:20. We attempted to fit 4 folds for each 64 candidates, resulting in 256 fits. The best model was selected from a range of hyper-parameters by using gridsearchCV. It's score on training data was 58.1 percent and on test data was 54 percent.

2.5 Observations

With an accuracy of 0.6, precision of 0.584, recall of 0.849, and an F1 score of 0.692, it's evident that the Adaboost Classifier, using DecisionTreeClassifier as its base model, performs reasonably well. The precision indicates that 58.4% of the predicted positive cases were correct, while the recall indicates that the model captured 84.9% of the actual positive cases. The F1 score, which balances precision and recall, sits at 0.692, suggesting a decent overall performance. However, the accuracy is relatively low, indicating that the model may struggle with correctly classifying a significant portion of the data. This could be due to a class imbalance issue, noisy data, or inadequacies in the model's hyper-parameters. Overall, while the model demonstrates promising performance in terms of recall and F1 score, there is room for improvement in accuracy and precision.

With an accuracy of 0.54, precision of 0.537, recall of 0.953, and an F1 score of 0.687, the RandomForest model demonstrates a mixed performance. The precision indicates that around 53.7% of the predicted positive cases were correct, while the recall indicates that the model captured a high percentage (95.3%) of the actual positive cases. The F1 score, which balances precision and recall, is 0.687, suggesting a fair overall performance. However, the accuracy is relatively low, indicating that the model struggles with correctly classifying a substantial portion of the data. This could suggest issues such as overfitting or class imbalance. While the model shows a high recall rate, indicating it effectively identifies positive cases, the precision is relatively low, suggesting a higher rate of false positives. Overall, while the model demonstrates promising performance in terms of recall and F1 score, there is room for improvement in accuracy and precision.

3 Super Market

3.1 Exploratory Data Analysis

The supermarket dataset has 1000 entries, in which each entry has 11 features about an invoice. There features are invoice id, branch of supermarket, customer type, gender of customer, category of product purchased, unit price of product purchased, number of units of the products purchased, tax amount, total amount on the invoice, mode of payment and rating submitted by the customer. Firstly, we explored dimensions of dataset, data types in which the features were stored and mean, range of values present for each feature. The dataset did not had any null values on any row. Correlation Heat-maps helped us understand correlations between all features. Then we performed one hot encoding on the categorical columns in the dataset. Further, we scaled the dataset using standard scaler, this ensured that the model will not be distorted by features having higher magnitudes.

3.2 Outlier Detection and Removal

We visualized distribution of features in the data by using box-plot, dis-plot and violon-plot. Then, the dataset was scanned using DBScan to check for presence of any outliers. The supermarket dataset had 8 outliers which were removed by using the IQR(inter-quartile range) method.

3.3 Train Test Splits

The dataset was split in ratio of 80:20, 80 percent of dataset was used for training and 20 percent was used for testing.

3.4 Models Implemented

3.4.1 Gender Detection

We initially attempted to train the decision tree classifier for the Gender Detection. The accuracy on training data was 57 percent and on test data was 51.3 percent. Further on attempting hyper-parameter tuning by using a range of maximum depths, minimum samples per leaf and selection criteria, the best model was using gini criterion, with a depth of 4 and minimum 38 samples per leaf. This resulted in increasing accuracy on training data to 58 percent and on testing data to 55.77 percent.

We then attempted to fit a random forest classifier. It resulted in an accuracy on 78 percent on training data, while the accuracy on test data dropped to 51.2 percent. Further, we attempted to optimize the classifier by using a range of hyper-parameters in the parameter grid. Best model from the grid resulted in 74.6 percent accuracy on training data and 54.77 percent accuracy on test data.

3.4.2 Rating Prediction

We attempted to train a linear regression in the data. The r2 score of the model on training data was 0.0096 and on test data was -0.01. Then we attempted to train the data to minimize ridge, lasso errors. Finally, we attempted to minimize lasso error with polynomial order features of degree atmost 2. This resulted in reducing mean absolute error to 1.40 in training data

Then, we attempted to train a decision tree regressor on the dataset. This resulted in a R2 score of 0.009 on training dataset and -0.0033 on test dataset. Then we performed hyper-parameter tuning on a range of parameters, by using the Grid Search on the grid of parameters. Finally, we got a best model in the grid which had a R2 score of 0 on training data and -0.00065 on test data.

3.5 Observations

3.5.1 Gender Prediction

3.5.1.1 Using Decision Trees With an accuracy of 0.558, precision of 0.615, recall of 0.647, and an F1 score of 0.630, the Decision Tree model showcases moderate performance. The precision indicates that around 61.5% of the predicted positive cases were correct, while the recall indicates that the model captured approximately 64.7% of the actual positive cases. The F1 score, which balances precision and recall, stands at 0.630, indicating a reasonable overall performance. Although the accuracy is above 0.5, suggesting the model performs better than random guessing, there's still room for improvement.

3.5.1.2 Using Random Forest

With an accuracy of 0.548, precision of 0.681, recall of 0.422, and an F1 score of 0.521, the Random Forest model displays a mixed performance. The precision indicates that approximately 68.1% of the predicted positive cases were correct, while the recall indicates that the model captured only about 42.2% of the actual positive cases. The F1 score, which balances precision and recall, stands at 0.521, suggesting a moderate overall performance. The model's accuracy is slightly above chance, but it's evident that there's room for improvement, especially in terms of recall. The high precision suggests that when the model predicts a positive case, it's often correct. However, the relatively low recall indicates that the model misses a significant portion of actual positive cases, resulting in an imbalanced performance.

3.5.2 Rating Prediction

3.5.2.1 Linear Regression

The best model we get here is with an R^2 score of 0.020, Mean Absolute Error (MAE) of 1.385, and Mean Squared Error (MSE) of 2.698, the Linear Regression with Feature Selection model demonstrates poor performance in predicting the target variable.

The low R2 score indicates that only about 2% of the variance in the target variable is explained by the model. Furthermore, the MAE of 1.385 suggests that, on average, the model's predictions deviate from the actual values by approximately 1.385 units. The MSE of 2.698 indicates that the model's predictions exhibit large errors, with squared deviations averaging around 2.698 units.

3.5.2.2 Decision Tree Regression

The best model we get is with an R2 score of approximately -0.001, a Mean Absolute Error (MAE) of 1.401, and a Mean Squared Error (MSE) of 2.754, the Decision Tree Regression model exhibits poor performance in predicting the target variable.

The negative R^2 score indicates that the model's predictions are no better than simply predicting the mean of the target variable. The MAE of 1.401 suggests that, on average, the model's predictions deviate from the actual values by approximately 1.401 units. Additionally, the MSE of 2.754 indicates that the model's predictions have large squared deviations from the actual values.

4 Conclusion

From the observations we can say that the relationships between the input features and the target variable might be complex, making it challenging for simpler models to capture them accurately.Also the selected features may not adequately represent the underlying relationships in the data. If the features are not informative or are noisy, it can lead to poor model performance, particularly in regression tasks. In conclusion, the poor performance of the models suggests that the data likely belongs to a domain with complex, nonlinear relationships, high noise, and possibly limited sample size.