Final Project Executive Summary

The Big Data Team

Myisha Chaudhry – 200591740

Kayleigh Habib – 200370580

Marcus Rilling – 200656520

Owen Bell – 200449530

Maheep Jain – 203386460

BU425: Business Analytics

Professor Michael Pavlin

November 29th, 2023

Executive Summary

The Telco Company (based in California) is handling the issue of customer churn as customers are choosing to discontinue the use of phone and internet services. This is occurring for a variety of reasons ultimately resulting in the company losing revenue for every customer that leaves. Aside from loss of business, customer churn has a negative impact as it gives the company a bad reputation. This could deter future customers from wanting to use the company's services as they know that past customers were not satisfied and chose to discontinue using said services. Customers leaving could also be the result of an underlying problem, such as having a bad product or being overpriced. The motivation for analyzing this business problem is to get a better understanding of the multiple factors that may be contributing to customer churn so that the company can take appropriate action.

The data for the customer churn case comes from IBM and consists of over 7,000 records and 33 columns, including customer characteristics as well as the different types of services and contracts the company offers. The "Churn Label" column is identified as the target variable, with "Yes" indicating the customer leaving. A bar graph indicates that this is very imbalanced, with 27% of the customers leaving. The "Total Charges" column is blank for a small number of records, and there are a few data fields that do not provide any predictive power, so these are removed from the analysis.

Several charts are created to explore the relationship between the target variable and the features, to understand the trends in the data, while a correlation plot is used to identify any multicollinearity among the predictors. The data exploration indicates that customer retention varies by contract type and payment method. In addition, customers who have been with the company for a longer time are less likely to leave, indicating customer loyalty. The predictive models used will help us understand the strengths and importance of these observations.

As this is a binary target response i.e., either the customer stays or leaves, classification models will be used to predict the outcome. The two machine-learning algorithms to focus on are Logistic Regression and Random Forests.

Logistic Regression is chosen due to its high interpretability and simplicity, as well as the fact that the target variable is binary. Using the predictors, the model will produce a probability between 0 or 1, indicating whether a customer is likely to leave. Random Forest is more complicated but produces accurate results, especially on imbalanced datasets. This model will tell us if a customer decides to leave or stay based on aggregating predictions from multiple decision trees. The model also does an excellent job of reducing variance and can be scaled easily for future projects.

The data is split into 3 sets: 80% of the records are used for training the models, 10% are used for testing and tuning the parameters while the final 10% is used for validating and comparing models. An over-sampling technique is applied to training data due to the target imbalance.

Models	Accuracy	Sensitivity	F1 Score	AUC
Logistic Regression	78.7%	77.6%	64.8%	0.870
Random Forest	75.5%	86.9%	64.2%	0.872

Both models produce similar AUC values, indicating they are equally viable to use in distinguishing whether a customer will churn or not. Comparing other metrics, the logistic regression has a higher accuracy score which represents the ability to correctly classify the customers. Reducing the volume of false negatives is important in this case as the company will be more concerned about misclassifying customers who are likely to leave. The random forest model with its higher sensitivity rate performs better by identifying a higher portion on these customers. With this information, the company will be able to focus its efforts and will miss only a small portion of customers who are likely to churn.

Both models indicate some of the important features in predicting customer churn include the Contract Length, Payment Method, and whether the customer has any Dependents. Using this information, the company can better tailor certain promotions to target these important features to limit customer churn and thus improve their profitability.

To finalize the results of this analysis it is recommended to use the machine learning models to enhance the telecom company's churn prediction to help in targeting preventative actions.