Customer Churn Analysis

Max Kutschinski

2022-10-26

Contents

Introduction 1

Exploratory Data Analysis 2

Modeling 6

Results 9

Conclusion 10

References 11

Introduction

In the telecommunications industry customers frequently switch from one service provider to the next. This makes it a highly competitive industry with an average annual churn rate of 22% (Liibert, 2021). Telco, a company providing home internet and phone services, is interested in testing a new strategy for increasing customer retention.

The company has determined that focusing on customer retention rather than acquisition best aligns with their long-term goal of increasing profits. According to their research, acquiring a new customer can cost around 5 times more than retaining an existing customer. Furthermore, even a small increase in customer retention can lead to a large increase in profits. Telco has found that customers are likely to spend more with companies they have already done business with. Furthermore, repeat customers are more likely to refer others, which will support long-term growth. This analysis focuses on predicting customer churn. By identifying customers who are at risk of churning, Telco can target these customers with incentives to stay and increase retention. The company is interested in a predictive model, as well as understanding important variables affecting customer churn. The company is also considering to send out a 20% discount code to customers who are at risk of churning and is inquiring about the profitability of the proposed incentive program.

Exploratory Data Analysis

The data set contains 7043 observations and 33 features and is made available on Kaggle. Each observation corresponds to a different customer, whereas the features relate to demographic information of the customers, such as gender, age, and location, as well as the types of services purchased and their cost. Customers are labeled as churned if they have left the business within the last month. Python is used to analyze the data and compile the report.

Figure 1 shows that Telco customers are located in California and clustered around big cities such as Los Angeles, San Francisco, and San Jose. There also doesn't seem to be any apparent relationship between churn rate and customer location.

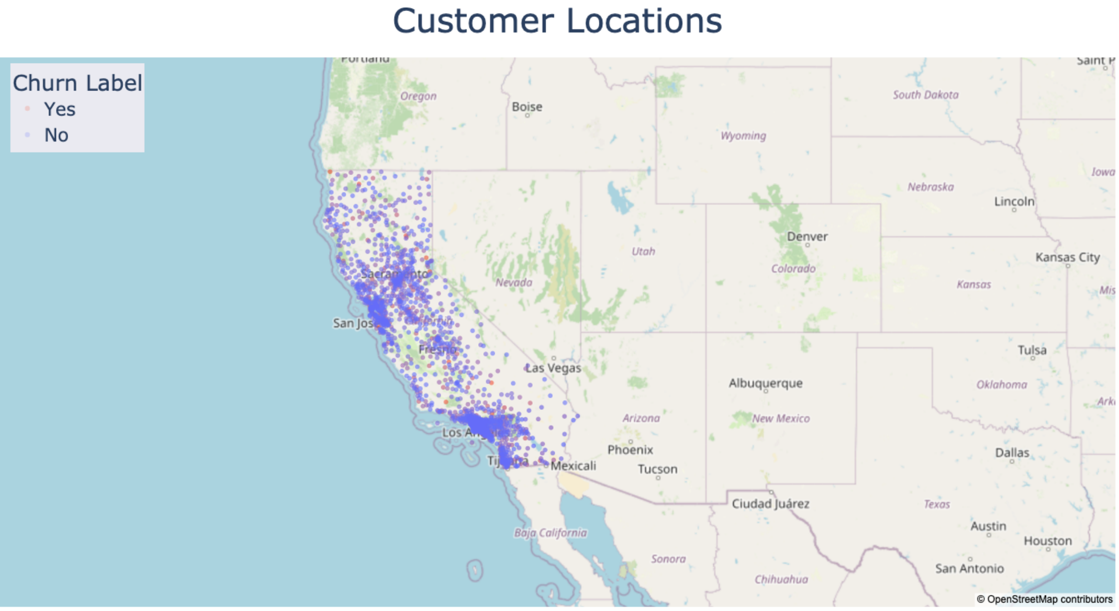


Figure 1: Customers are located in California and clustered around large metropolitans. There does not appear to be a relationship between churn rate and customer location.

As seen in Figure 2, the current churn rate is around 25%. In other words, approximately 1 out of 4 customers ends up cancelling their business with the company.One way of gauging the incurred loss due to churned customers is by estimating their overall value to the company. The customer lifetime value (CLTV) estimates a customer's value and is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn since they are the most profitable and it costs less to keep existing customers than it does to acquire new ones.

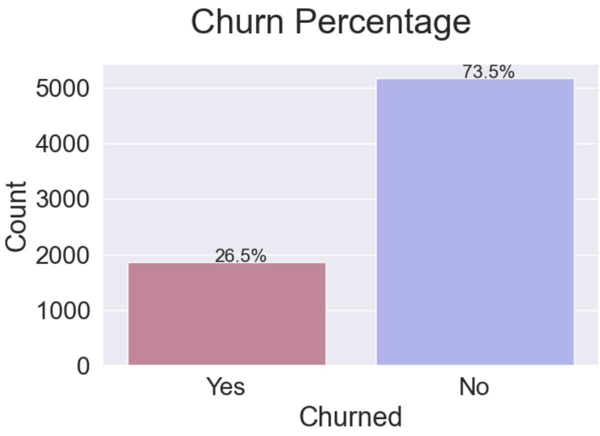


Figure 2: The current churn rate is around 25%.

Figure 3 displays how CLTV is distributed among churned and and non-churned customers. It also takes into account the total number of months that a customer has been with the company. On average, churned customers have a lower life-time value to the company. Furthermore, long-term customers have a higher value and are less likely to churn.

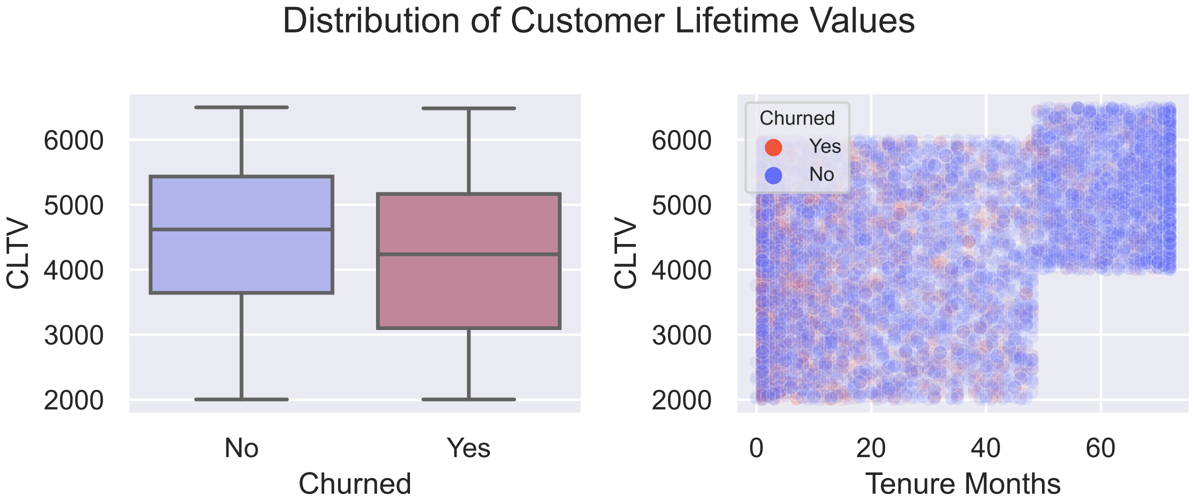


Figure 3: On average, the estimated CLTV is lower in churned customers. Long-term customers have a higher value and are less likely to churn.

The data set also contains information on the total monthly charges for each customer. This can be used to calculate the total revenue that is lost due to churned customers. Overall, churned customers constitute around $140,000 in lost revenue, which translates to 30% of total revenue. Figure 4 summarizes the distribution of monthly charges across churned and non-churned customers and indicates that higher monthly charges seem to correlate with higher churn rates.

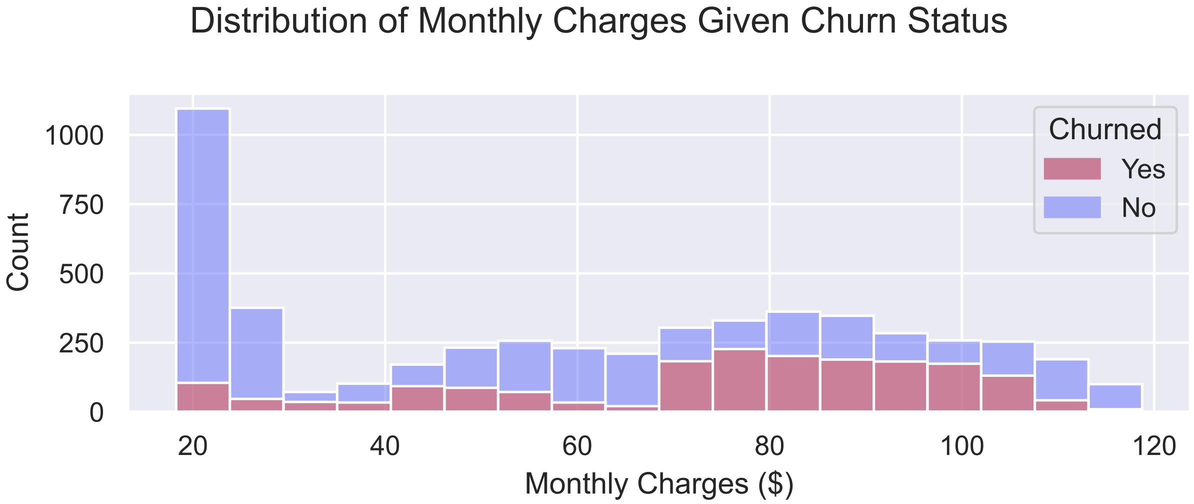


Figure 4: Higher monthly charges are correlated with higher churn rates

Total monthly charges are calculated based on the different kinds of services that a customer is subscribed to. These services include phone service, internet service, multiple lines, online security, online backup, device protection, and tech support. Figure 5 summarizes the popularity of these services by displaying the number of customers subscribed to them, as well as the churn rate within each service. The majority of customers are subscribed to phone and internet services, whereas tech support and online security are the least popular services. Furthermore, customers that bought the internet service have a high churn rate and customers that are subscribed to less popular services seem to be less likely to churn.

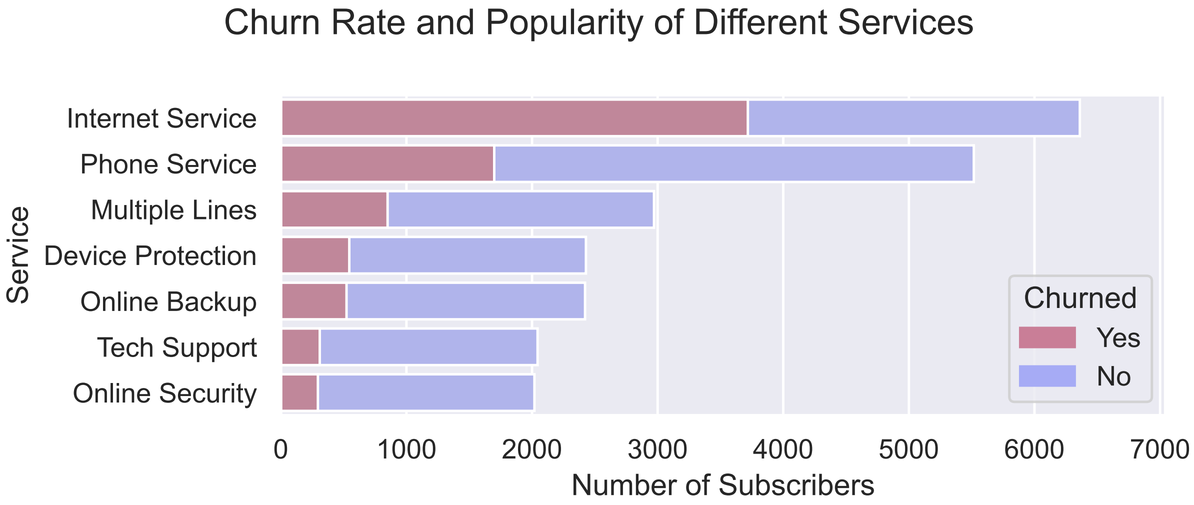


Figure 5: The majority of customers are subscribed to phone and internet services. Tech support and online security are the least popular services. Churn rate decreases with popularity.

Customers can choose between three different types of contracts: month-to-month, one year, and two year. In addition, they have the option to opt in to paperless billing and choose between different payment methods. Churn rates of different payment related variables are displayed in Figure 6.

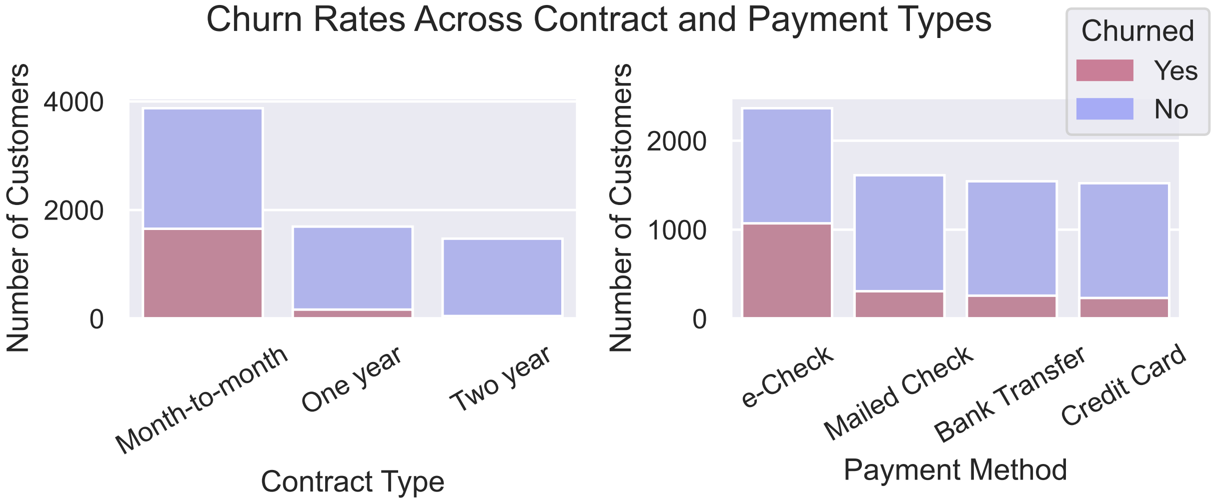


Figure 6: Long-term contracts have very low churn rates, but are also less popular. Paying with electronic check is the most common and has the highest churn rate.

Month-to-month contracts are the most popular option and are associated with the highest churn rates. On the other hand, long-term contracts are less popular but have a very low churn rate. Most customers pay via electronic check, which has a higher churn rate than other payment methods.

Demographic information on customers include whether they are a senior citizen, have dependents, as well as their gender and if they have a partner. Figure 7 plots churn rates after taking demographic information into account.

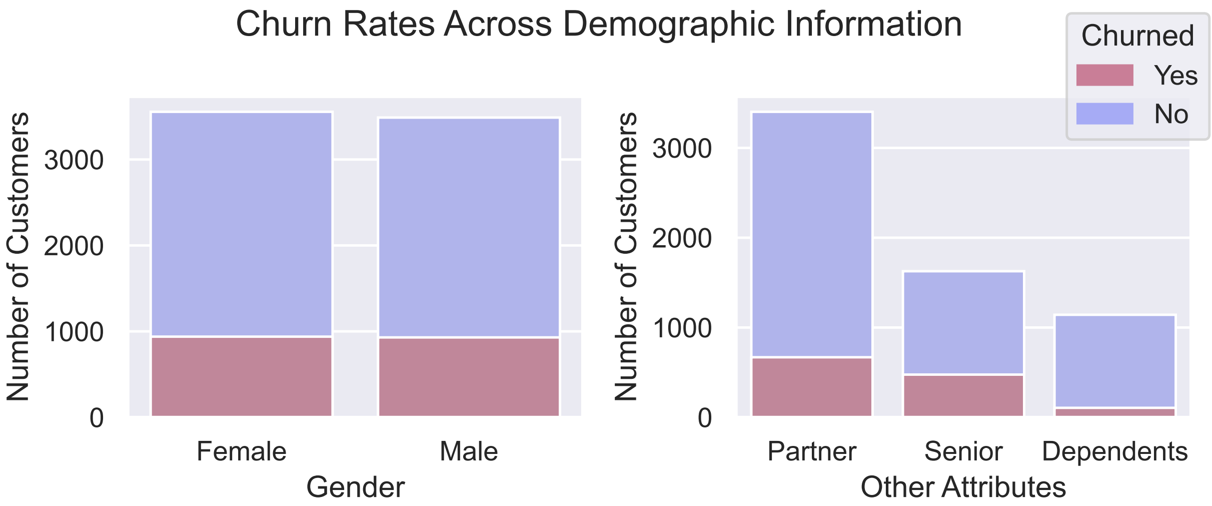


Figure 7: Churn rates do not differ significantly between genders. Customers with dependents have a comparatively low churn rate.

There are a roughly equal number of male and female customers and churn rates do not differ significantly between the two genders. Furthermore, customers with dependents have a relatively low churn rate.

A simple way of gaining insight into the reason why customers are churning is by asking them directly. Telco records the responses to a survey that asks customers who are cancelling their services about the specific reason for leaving.

Figure 8 visualizes the most common words from these customer surveys by generating a word cloud, where a larger font size corresponds to a higher frequency of the word. Overall, the most common reported churn reason relates to competitors offering better services. Other words that stand out, such as support and attitude, suggest that negative customer support experiences are another frequent churn reason.

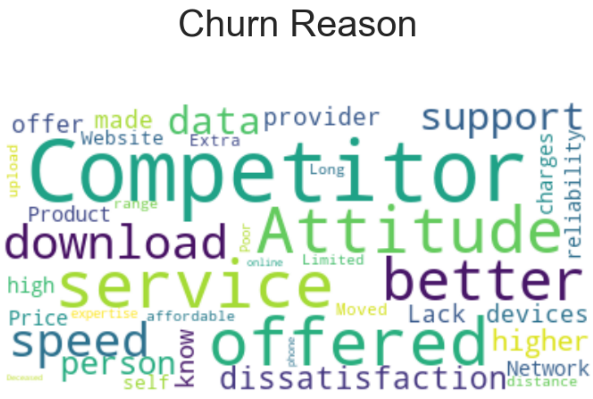


Figure 8: Most customers churn due to competitors and negative customer support experiences.

The heatmap in Figure 9 describes the correlation among the four quantitative features in the data set. Since total charges are calculated as the product of monthly charges and tenure, it is no surprise that they are highly correlated.

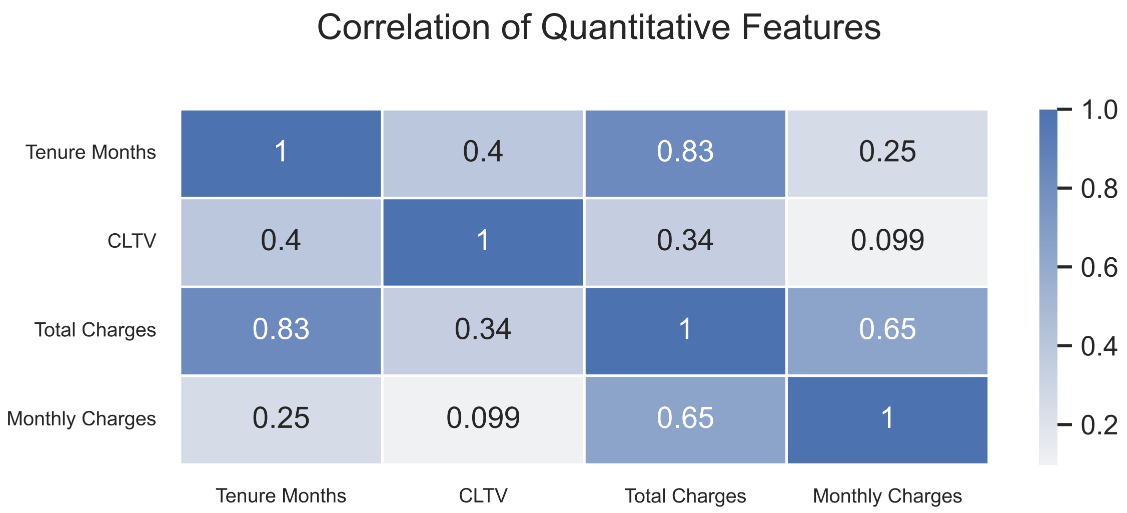


Figure 9: Total charges are highly correlated with tenure and monthly charges.

Modeling

Total Charges is the only feature with missing values. These observations correspond to customers that have a tenure of less than a month. Since churn values are established for customers that churned within the last month, the best way to deal with these observations is to drop them. Note that since the entire feature is highly correlated with Tenure Months and Monthly Charges, it is not very informative to begin with and can even negatively affect some of the models. Therefore, the entire feature is excluded from the analysis.

Most machine learning models don't natively support non-numeric data. Hence, all categorical features are manually transformed into dummy variables via one-hot-encoding prior to being used as input for modeling. In addition, uninformative features such as Customer ID, Country, and State are dropped. This results in a new data set with dimensions 7032 x 46.

Another thing to consider before training models is that there exists an unequal proportion of churned and non-churned customers in the data set, which will negatively impact prediction performance of the minority class. Synthetic Minority Over-Sampling Technique (SMOTE) is used to address this issue by creating a more balanced training data set.

Different classification models are used to predict churning customers. These include tree ensembles such as Random Forests and XGBoost, as well as other parametric models such as Logistic Regression and Logistic Elastic Net. All models are trained on a training set and evaluated on an independent test set. Classification results are summarized in figures 10 to 13. Each figure is comprised of a table of performance metrics and a confusion matrix, where 1 represents churned customers and 0 represents non-churned customers.

Figure 10 shows that the Random Forests model has high overall Accuracy but does not do very well at predicting the minority class (customer churn).

Random Forests

Metric	Score
Accuracy	0.798
Precision	0.653
Recall	0.513
F1-Beta	0.595

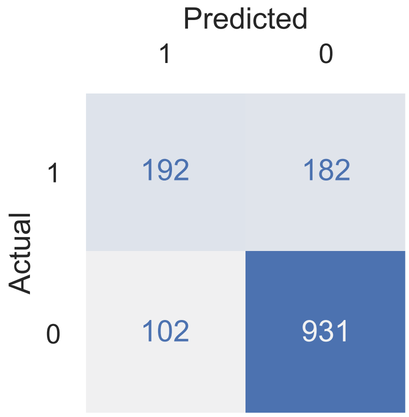


Figure 10: Random Forests: High Accuracy, low Recall.

Figure 11 summarizes the XGBoost results, which has lower Accuracy than the Random Forests model but performs much better at predicting customer churn.

XGBoost

Metric	Score
Accuracy	0.66
Precision	0.432
Recall	0.89
F1-Beta	0.531

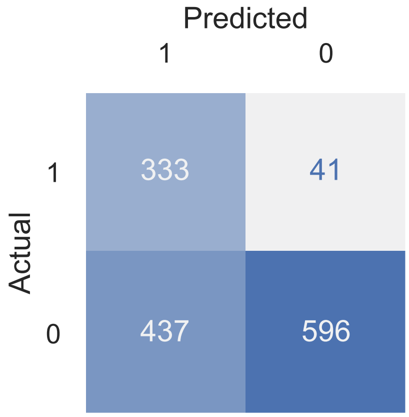
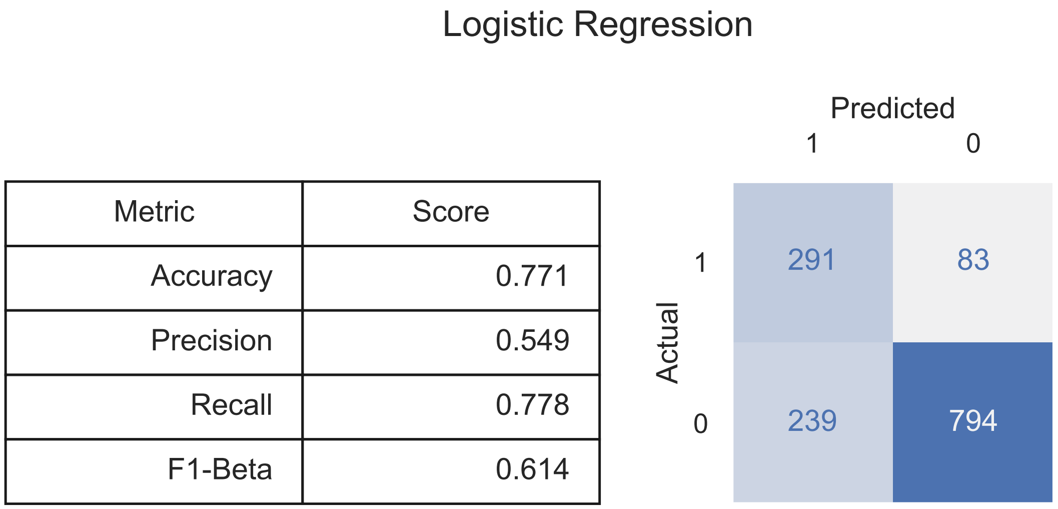


Figure 11: XGBoost: Low Accuracy, high Recall.

Figure 12 displays the Logistic Regression prediction results with Accuracy and Recall scores falling in between the two tree-based models.

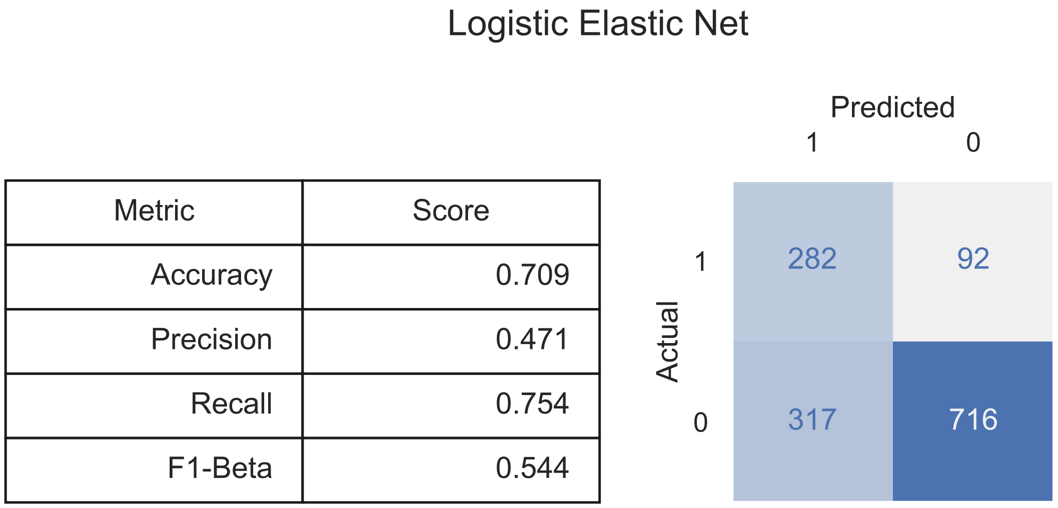

Logistic Regression

Metric	Score
Accuracy	0.771
Precision	0.549
Recall	0.778
F1-Beta	0.614

		Predicted	
		1	0
Actual	1	291	83
	0	239	794

Figure 12: Logistic Regression: Medium Accuracy, medium Recall.

The Logistic Elastic Net predictions are summarized in Figure 13. This model performes slightly worse than Logistic Regression across all metrics.


Logistic Elastic Net

Metric	Score
Accuracy	0.709
Precision	0.471
Recall	0.754
F1-Beta	0.544

		Predicted	
		1	0
Actual	1	282	92
	0	317	716

Figure 13: Logistic Elastic Net: Medium Accuracy, medium Recall.

Permutation feature importance is a technique that is used to measure the importance of features for a particular model. It is defined to be the decrease in a model score when a single feature value is randomly shuffled. If a feature is highly important then the prediction score should decrease substantially after permuting the values for that feature. Figure 14 plots the permutation feature importance of the XGBoost model and indicates that contract type and tenure are highly important predictors of customer churn. Monthly charges, internet service, and dependents are also among the most important features.

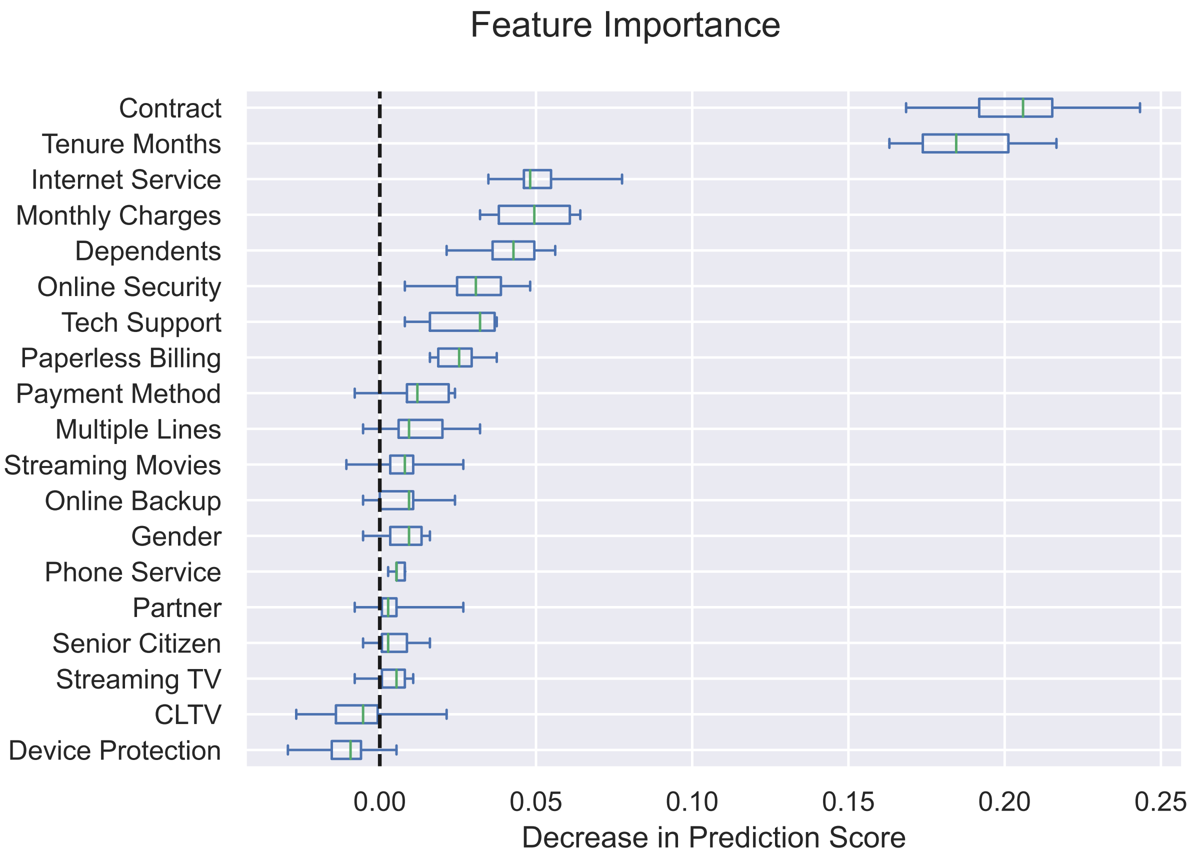


Figure 14: Contract type and tenure are highly important predictors of customer churn. Monthly charges, internet service, and dependents are also among the most important features.

Results

All models face a trade-off between Type I and Type II error and one must carefully consider this trade-off, along with the business implications, when deciding on an evaluation metric. A Type I error occurs when a customer is falsely predicted to churn. On the other hand, a Type II error occurs when a customer is falsely predicted not to churn. For Telco this means that a Type II error would result in targeting non-churning customers with promotions directed at churning customers. Although this is certainly less than ideal, it is arguably more desirable than failing to identify churning customers. Thus, this analysis favors models with low Type I error as opposed to low Type II error.

In terms of performance metrics, this means that metrics such as Recall and the F1-Beta score are preferred over Accuracy. Recall (sometimes referred to as Sensitivity or True Positive Rate) measures the fraction of actual churned customers that were correctly classified. The F1-Beta score uses a weighted approach to strike a balance between Recall and Precision, where Precision measures the fraction of churned predictions that are actual churned customers.

Precision is important in the context of discount codes because a low Precision score translates to more discount codes being handed out to non-churning customers, which would negatively impact the campaigns profitability. The best model maximizes Precision and Recall and yields the highest Return on Investment (ROI). The ROI can be calculated as $ROI = \frac{Revenue - Cost}{Cost} * 100$ where $Revenue = P(Success) * (1 - Discount) * Monthly\_Cost$ and $Cost = Discount * Monthly\_Cost$. Assuming a 40% success rate and a

20% discount, the ROI for the Logistic Regression Model is 60%, making it the most profitable model.

Conclusion

The primary focus of this analysis was on developing a predictive model for identifying customers who are at risk of churning. This model was then used to evaluate the profitability of an incentive program to increase customer retention. It was shown that a targeted 20% discount would result in a 60% ROI under the best performing model. Furthermore, the analysis has revealed which variables have the strongest effect on customer churn. New customers and customers with high monthly charges are most at risk of churning. In addition, customers with an internet subscription are churning at higher rates than other services. The two main churn reasons reported by customers are related to competitors offering better services and dissatisfaction with customer support.

Future research should be done to investigate self-reported churn reasons. For example, there are a variety of variables related to customer support interactions, such as chat logs, wait time, and the number of exchanges, that could be collected and added to the data set. This would allow Telco to address shortcomings in the services they provide, ensure a better overall customer experience, and ultimately increase customer retention.

References

Liibert, Katheriin. "Essentials of Customer Churn and Retention | Smartlook Blog." Smartlook Blog, 26 Apr. 2021, https://tinyurl.com/yy2sp33t