Customer Churn Analysis

Max Kutschinski

2022-10-23

Contents

Introduction

Telco, a company providing home internet and phone services, is interested in identifying customers who are at risk of churning. Customers can easily switch from one service provider to the next, making the telecommunications industry highly competitive with an average annual churn rate of 22%. 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. By identifying customers that are likely to churn, Telco can launch targeted business campaigns to this subset of customers in an effort to increase retention.

Exploratory Data Analysis

The data set contains 7043 observations and 33 features. 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. 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.

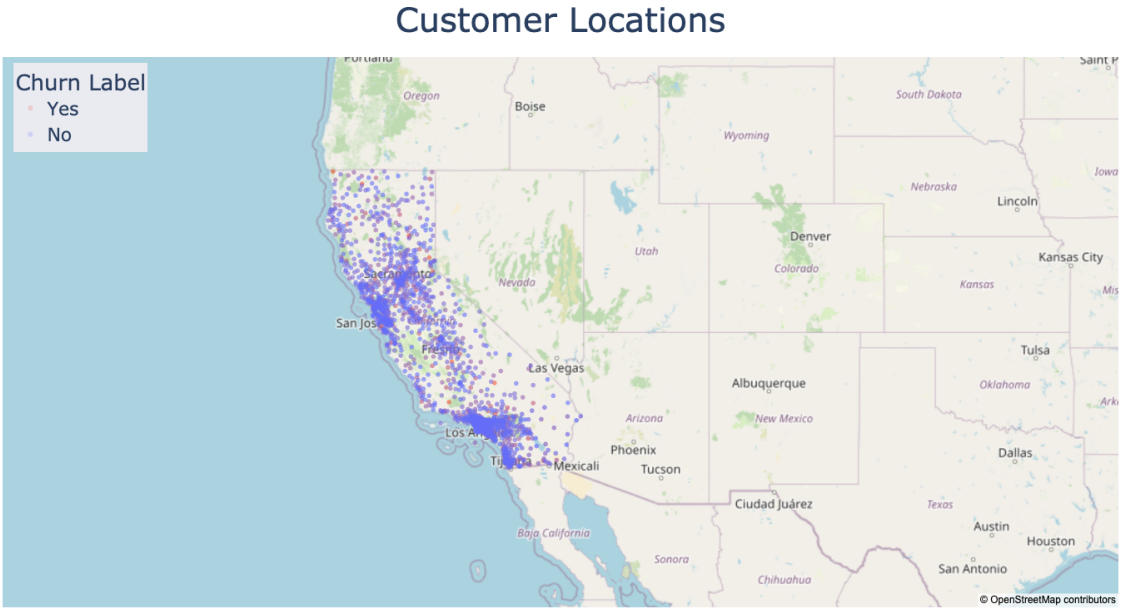


Figure 1: Customers are located in California and clustered around large metropolitans. There does not appear to be a relationship between churn rate and customer location.

The current churn rate is around 25%. In other words, approximately 1 out of 4 customers ended up cancelling their business with the company (Figure 2), which is quite high. 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 profitabland it costs less to keep existing customers than it does to acquire new ones.

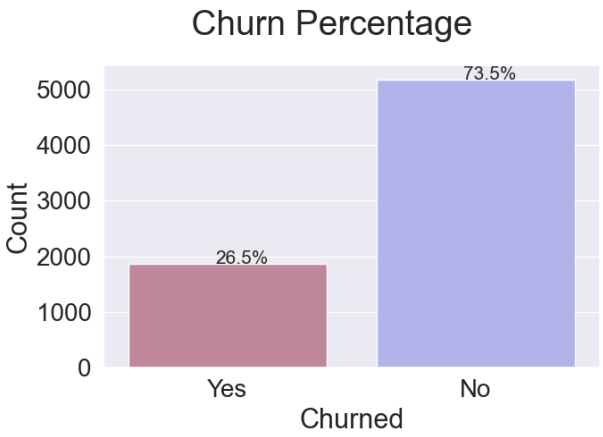


Figure 2: The current churn rate is around 25%.

Figure 3 displays how the CLTV is distributed among churned and and non-churned customers. It also takes into account total number of months that the 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.

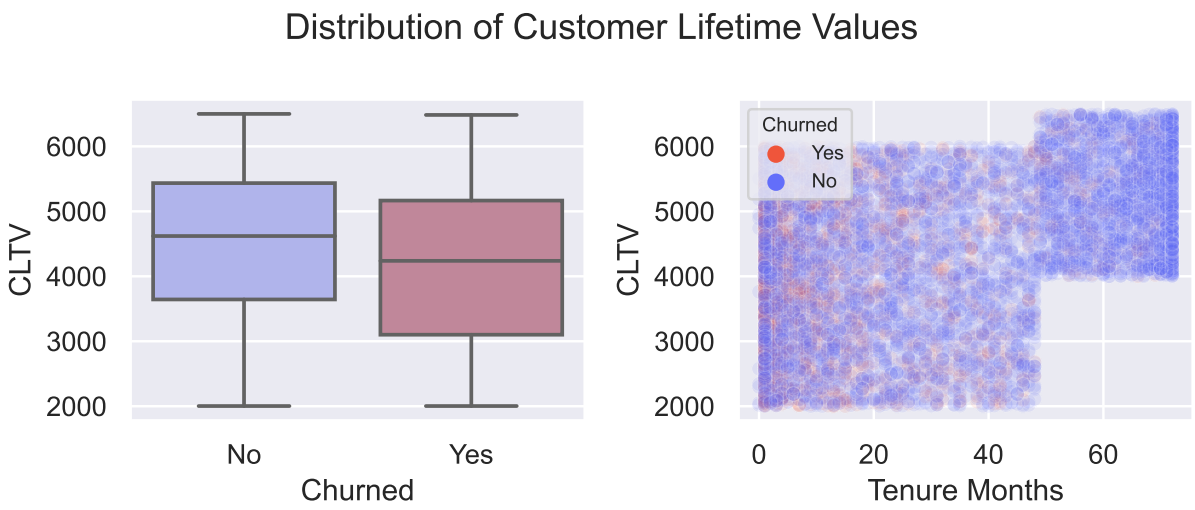


Figure 3: On average, the estimated CLTV is lower in churned customers. Long-term customers have a higher value and are less likely to churn.

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.

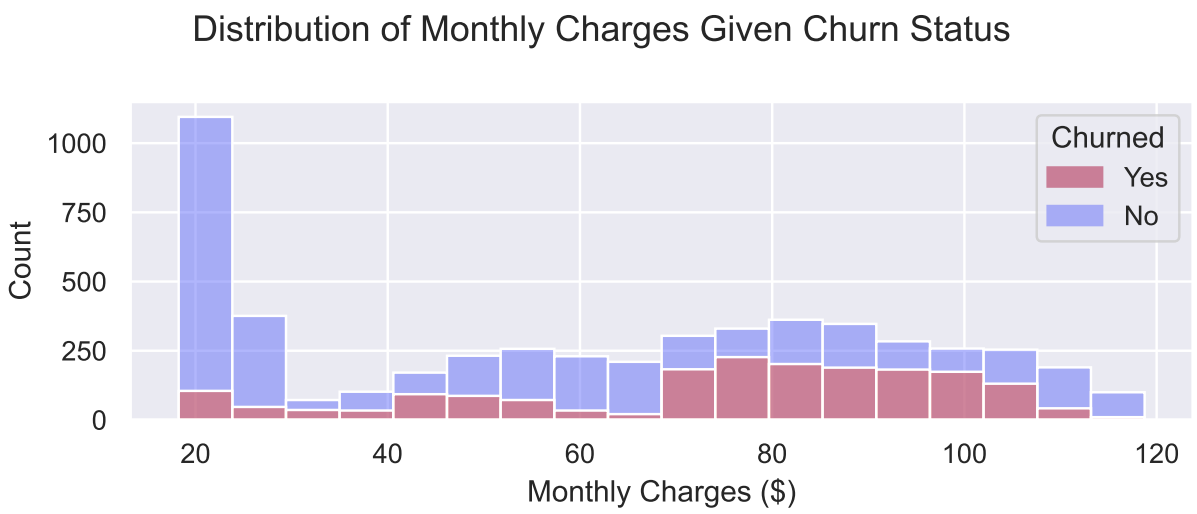


Figure 4: Higher monthly charges are correlated with higher churn rates

Total monthly charges are calculated based on different services that a customer is subscribed to. These services involve 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 are subscribed to less popular services seem to be less likely to churn.

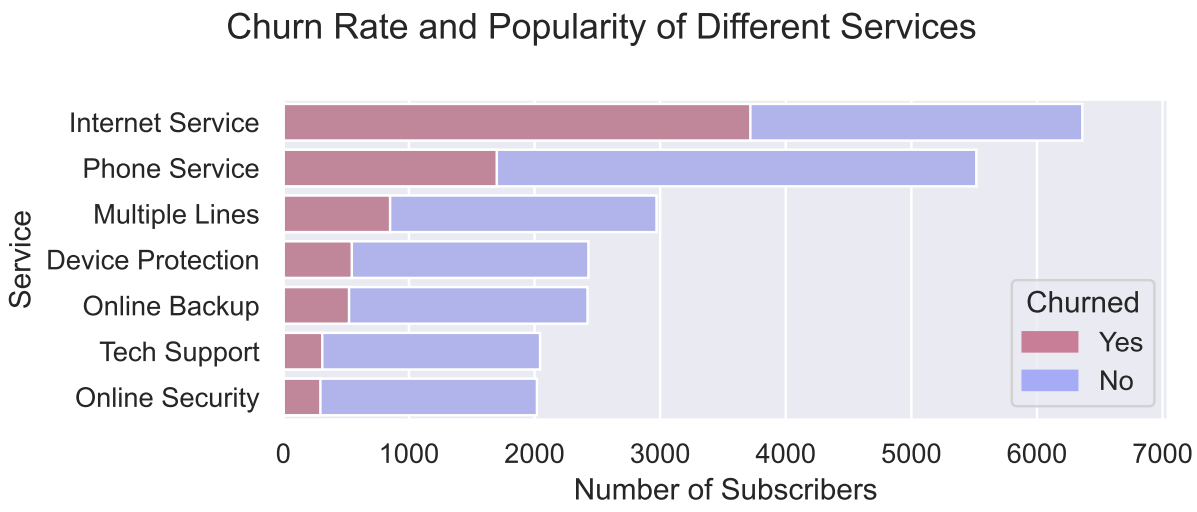


Figure 5: The majority of customers are subscribed to phone and internet services. Tech support and online security are the least popular services. Churn rate decreases with popularity.

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 choose between different payment methods. Churn rates among different payment related variables are displayed in Figure 6.

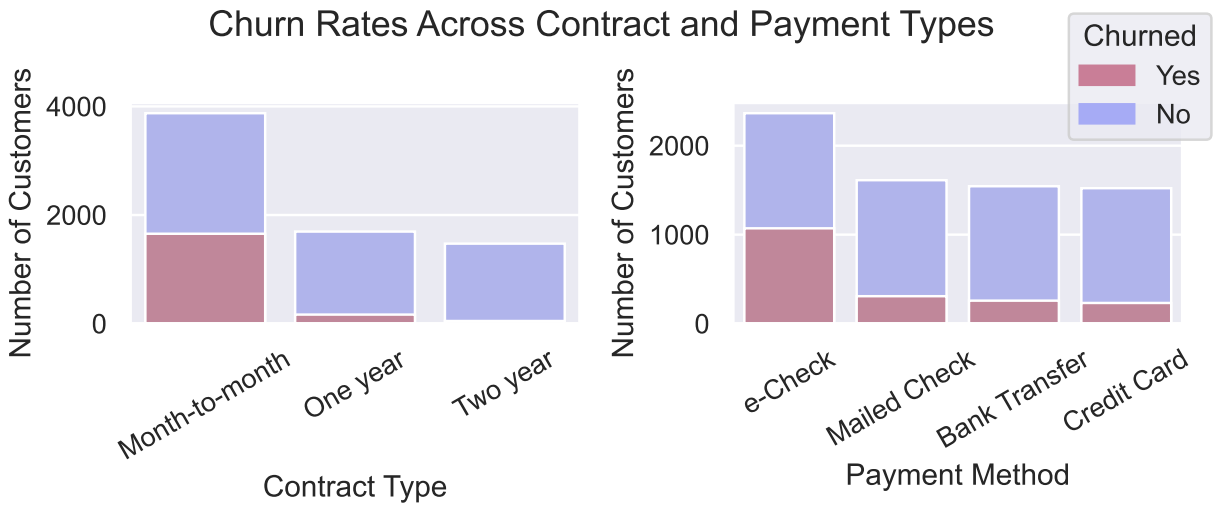


Figure 6: Long-term contracts have very low churn rates, but are also less popular. Paying with electronic check is the most common and has the highest churn rate.

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 the customers include whether they are a senior citizen, have dependents, or have a partner, as well as their gender. Figure 7 plots churn rates by taking into account demographic information.

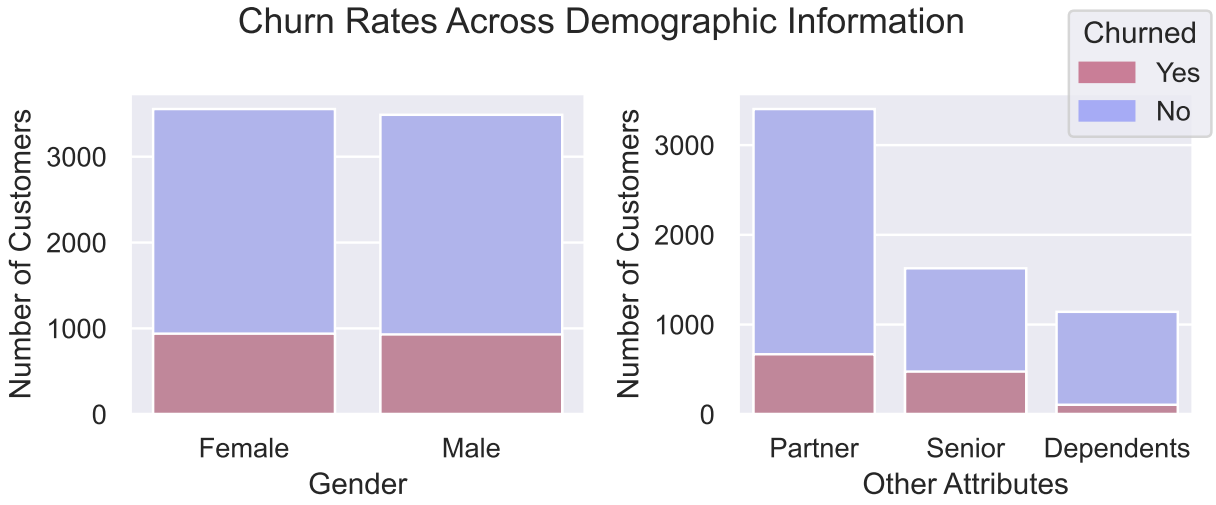


Figure 7: Churn rates do not differ between genders. Customers with dependents have a comparatively low churn rate.

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

A simple way of gaining insight into 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 churn reason can be attributed to competitors. Other words that stand out, such as support and attitude, suggest that negative customer support experiences are another frequent churn reason.

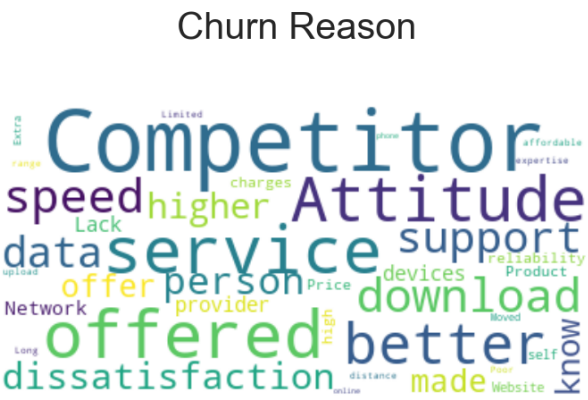


Figure 8: Most customers churn due to competitors and negative customer support experiences.

The "Total Charges" feature has 11 observations 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.

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

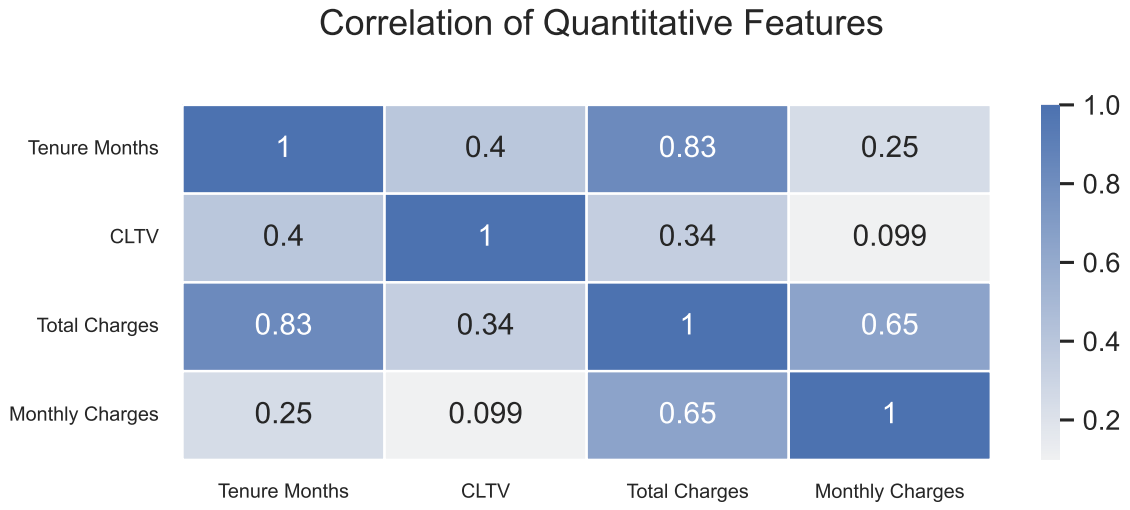


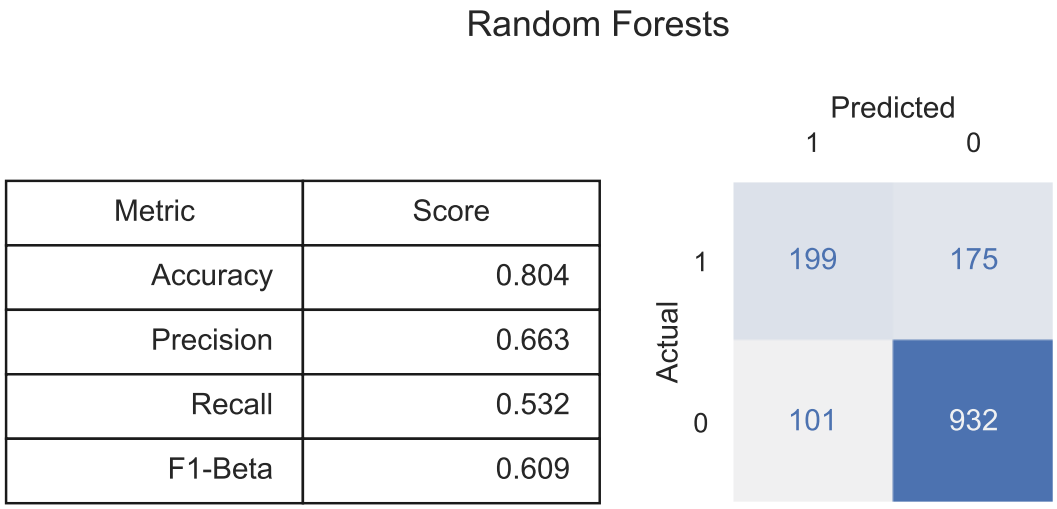
Figure 9: Total charges are highly correlated with tenure and monthly charges.

Modeling

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 frame with dimensions 7032 x 47.

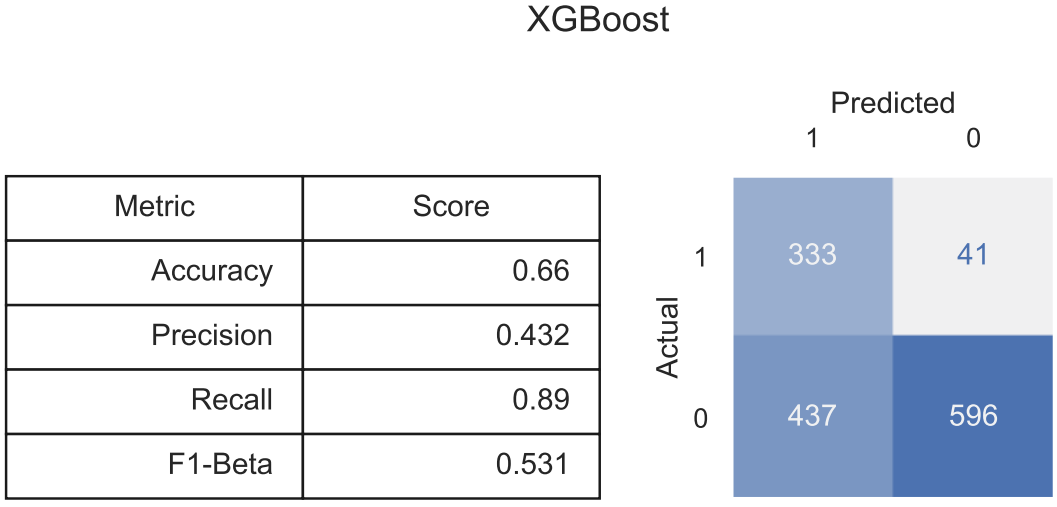
Another thing to consider before training models is that there exists an unequal proportion of churned and non-churned customers in the data set, which will negatively impact prediction performance of the minority class. Synthetic Minority Over-Sampling Technique (SMOTE) is used to address this issue and create a more balanced 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. Classification results are summarized in figures 10 to 13. Each figure is comprised of a table of performance metrics on and a confusion matrix, where 1 represents churned customers and 0 represents non-churned customers.



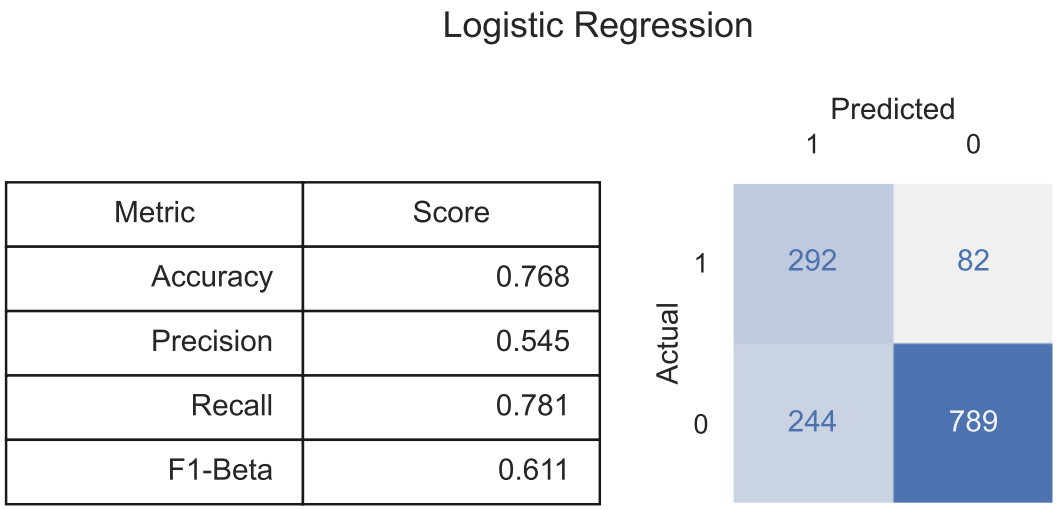
Metric	Score
Accuracy	0.804
Precision	0.663
Recall	0.532
F1-Beta	0.609

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



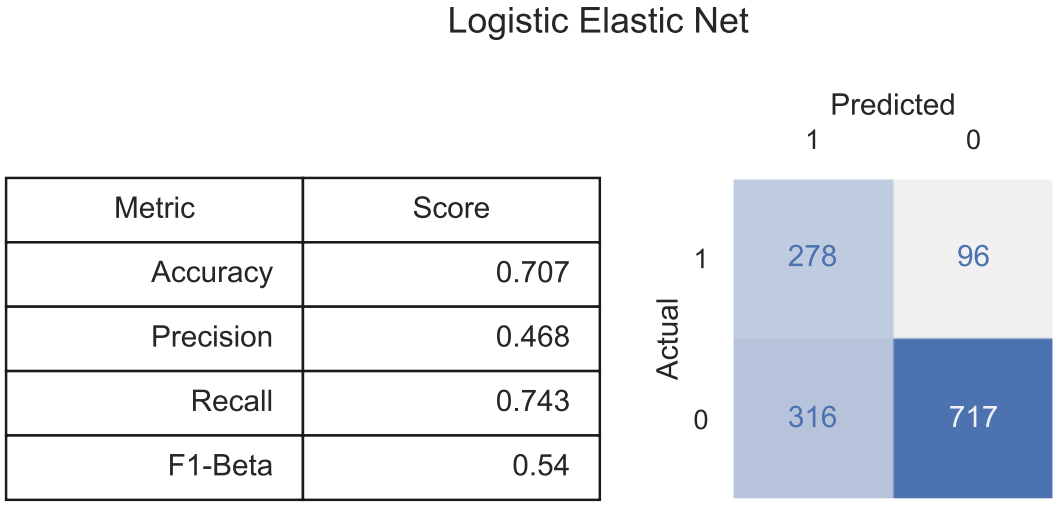
Metric	Score
Accuracy	0.66
Precision	0.432
Recall	0.89
F1-Beta	0.531

Figure 11: XGBoost: Low Accuracy, high Recall.



Metric	Score
Accuracy	0.768
Precision	0.545
Recall	0.781
F1-Beta	0.611

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



Metric	Score
Accuracy	0.707
Precision	0.468
Recall	0.743
F1-Beta	0.54

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

Results

All models face a trade-off between Type I and Type II error. It is important to carefully consider this trade-off and 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 campaigns directed at churning customers. Although this could have negative consequences in the long-run, it is arguably more desirable than failing to identify churning customers. Thus, models with low Type I error are preferred over models with low Type II error.

For this reason, performance 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.

Table X summarizes the model results and indicates that X algorithm performs the best when measured by Recall and F1-beta score.

Assuming that the target campaigns have a 50% effectiveness ....

```
library(dplyr)
library(kableExtra)
kable(py$df, digits=2)%>%
  kable_styling(font_size = 12, latex_options = "hold_position")%>%
  row_spec(0, bold = TRUE)%>%
  kableExtra::pack_rows("Baseline", 1,1 )%>%
  kableExtra::pack_rows("Top 3 SARIMA", 3,3 )
```

Conclusion

Summarized the findings of the report. How this will benefit people. Where future research could go.