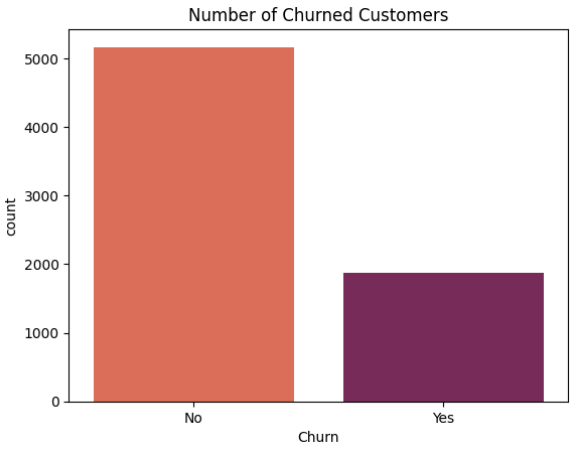
Telecommunication Customer Churn Prediction

Melissa Harrison

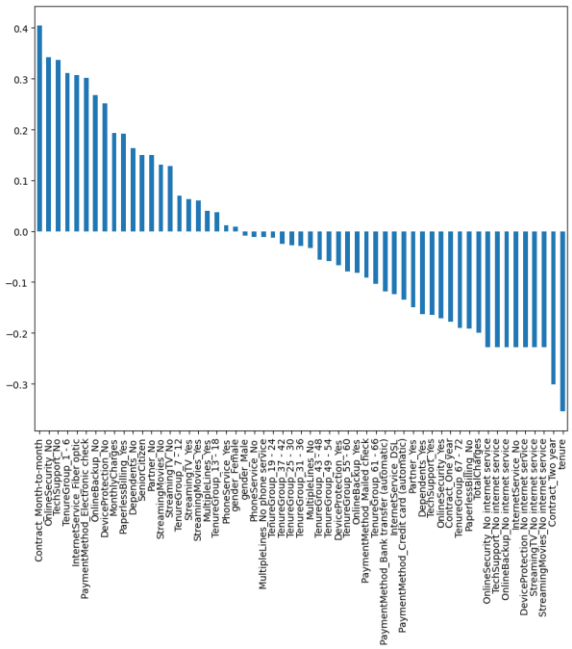
Customer churn is a critical part of businesses, especially in service industries, and in this case telecommunications. In this report, we'll go over the dataset and analyze it to discuss what the main factors that cause churn are. We'll also go over the models created to predict if a customer will churn.

First of all, let's make sure we understand what churn is in a business setting. Typically, customer churn is the percentage of customers a business loses over a period. Minimizing this churn is important because acquiring new customers is far more expensive than retaining current ones. In terms of telecommunication, this churn could look like a customer that cancelled their account to switch providers, maybe for better prices, features or service.

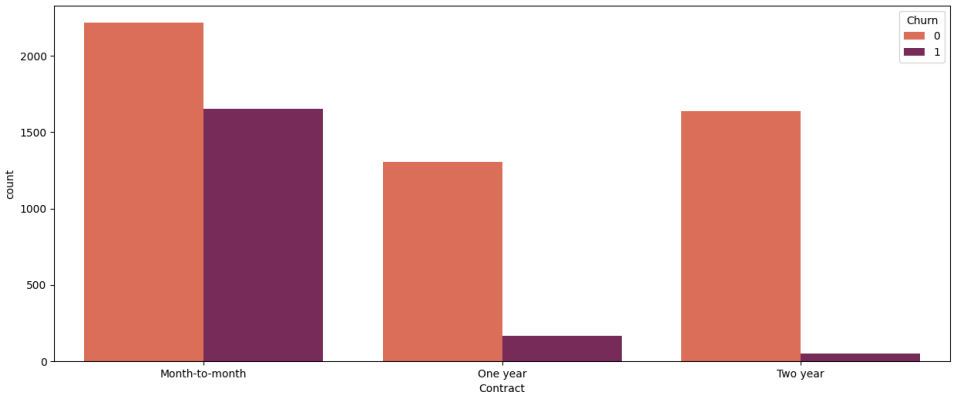
Since telecom companies tend to have a large customer base, it's hard and often impractical to personalize retention strategies, and this is where predictive modeling can help. Predictive modeling can be a large-scale solution to identify customers who are likely to leave before they actually do. This can allow a company to intervene with things like targeted offers or new contract plans aimed at a certain demographic and maximize the efficiency of customer retention strategies. Now, let's look at the dataset, and see what trends or patterns appear after analysis.

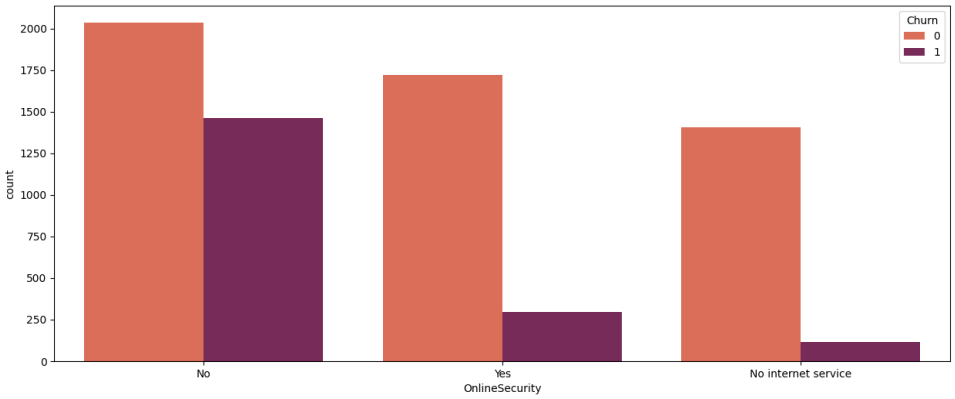


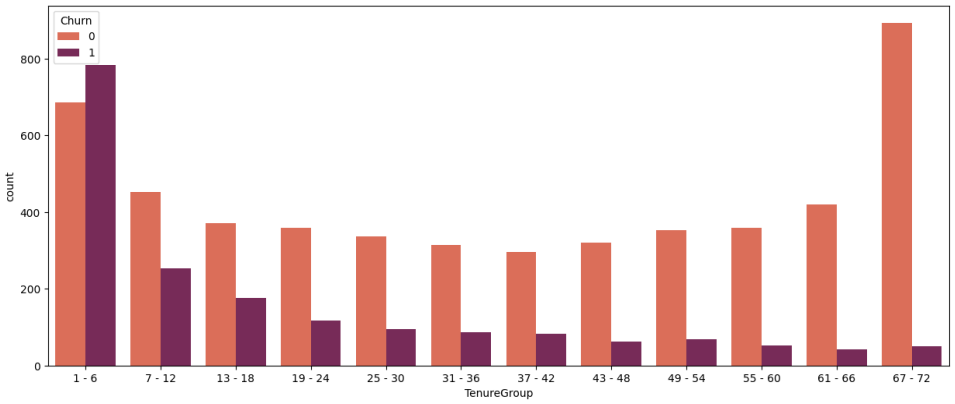
An initial breakdown shows us that there are a majority of the customers are being retained and not churning. But, a pretty sizable portion, over 26% of customers in our recorded data have churned. Now we also have to remember that the no-churn category can include everyone from customers that have been there for years, to people who just started.

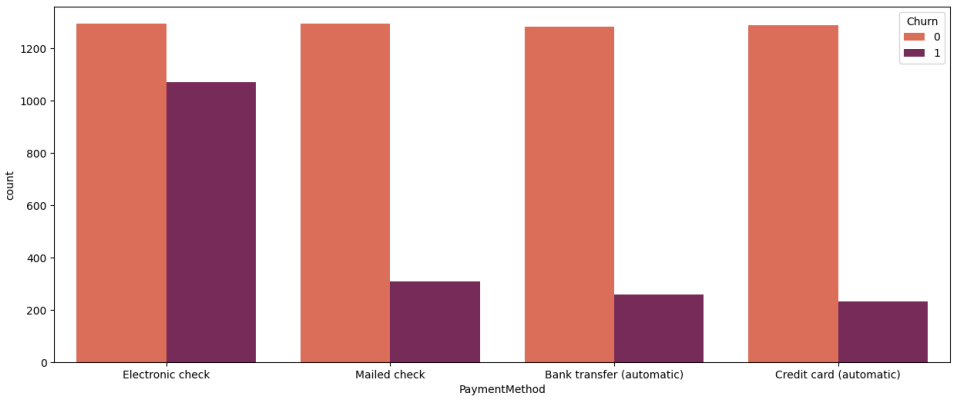


From this chart on correlation, we can clearly see what variables have the most impact on churn, and which tend to help prevent churn. Month to Month contracts, no online security, no tech support, a tenure of less than 6 months, and fiber optic internet service all positively correlate with churn. On the other side, for the customers with long term contracts over 2 years, subscriptions without internet, and with the longest tenures seem to be the least likely to churn. We can also see that some factors have little impact, like multiple lines or the gender of the customer. Now let's look at some charts that help visualize these trends.


Churn
0
1
count
2000
1500
1000
500
0
Month-to-month
One year
Two year
Contract

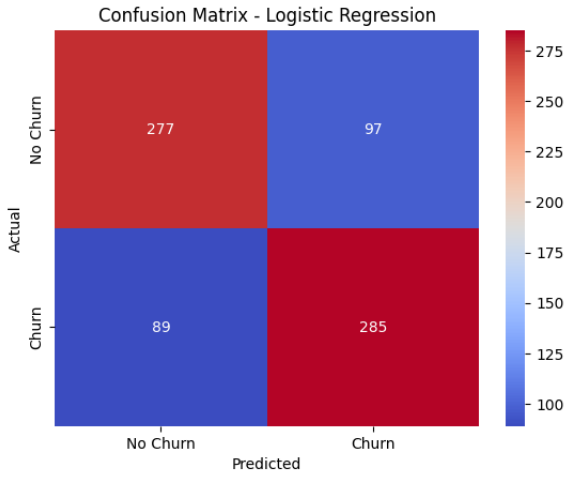

Churn
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
OnlineSecurity


Churn
0
1
count
800
600
400
200
0
1 - 6
7 - 12
13 - 18
19 - 24
25 - 30
31 - 36
37 - 42
43 - 48
49 - 54
55 - 60
61 - 66
67 - 72
TenureGroup

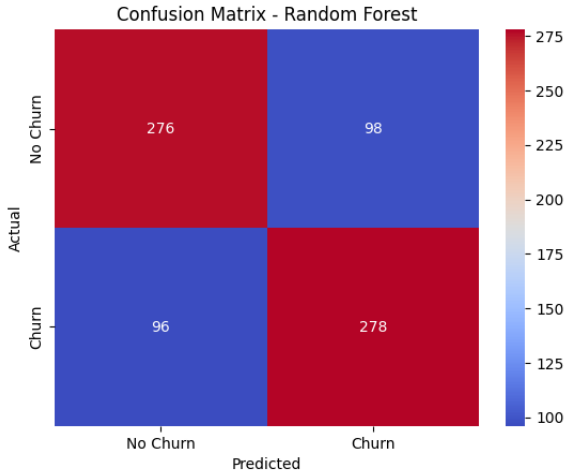

Churn
0
1
count
1200
1000
800
600
400
200
0
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

Now that we understand the data and its trends a bit more, it's time for modeling and predicting customer churn. We will go with regression models, as churn is a binary of either yes or no. However, how we balance the data is important, as we have a high majority of customers that don't churn, and that can skew our predictions to be biased.

The first attempt to balance the data was by under sampling, where we cut down the amount of non-churn customers to match the amount of those that did churn. However, we'll be able to see how this method has some flaws.



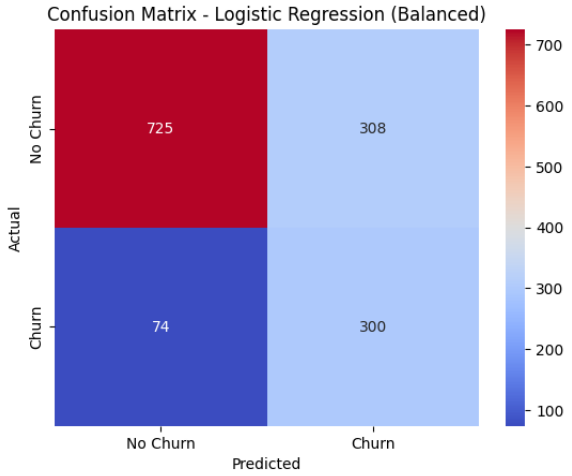
This linear regression model using under sampling has an accuracy of 75.1% and correctly identified 76% of churners with a balanced performance across both classes. This means it is doing a solid job of finding customers likely to churn while keeping false positives reasonably low. Its performance is consistent and reliable, making it a fairly strong candidate for churn detection.



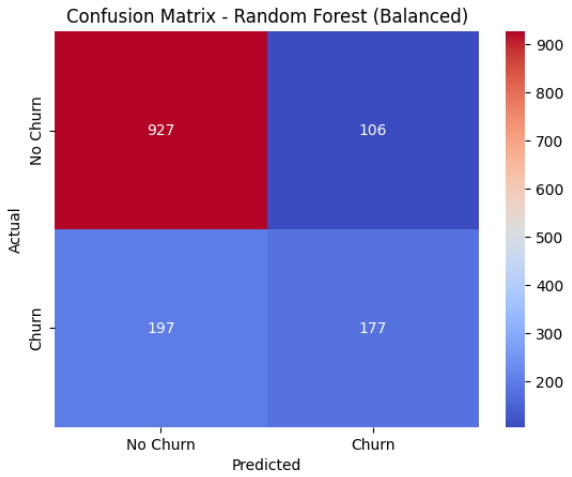
The under sampled Random Forest model had a slightly lower accuracy of 74%, with a recall of 74.3% for churners. It correctly identified 278 out of 374 churners, very close to the performance of Logistic Regression. However, the results suggest a slightly less balanced outcome, with a few more missed churners and a marginally lower overall score.

Now we can see where under sampling has some issues. By removing a large number of the non-churn class, the model loses a lot of information. This means our models are less accurate as it hasn't seen enough examples of churn, and it doesn't do as well on new testing data. So, let's try another way of balancing the data that doesn't lose a large portion of our data.

The second way we can balance the data is by balancing based off of class weights, where we can tell the model to pay more attention to the churn class. This means we don't under sample the data, and therefore with the full dataset are able to get an overall picture that's better at prediction.



This balanced Logistic Regression model had an accuracy of about 72.8%, but more importantly, it correctly identified 80% of customers who actually churned. This high recall means it was very effective at spotting customers likely to leave, even though it also incorrectly flagged some customers who actually didn't churn. This is acceptable in this churn scenario because it's far better to reach out or offer deals to a few customers that were going to stay anyways, than to miss a customer who is about to leave.



The balanced Random Forest model had a slightly higher accuracy at 78.5%, but it only caught 47% of actual churners. That means it missed more than half of the customers who were going to churn. While its predictions were a bit more precise, the low recall is far

less helpful in trying to prevent as much churn as we can, and overall this is a poor model to use.

This project focused on predicting customer churn in the telecommunications industry, because keeping existing customers is more valuable than acquiring new ones. After analyzing the data, we found several key factors that influence whether a customer is more likely to leave, including factors like short contract lengths, low tenure, and lack of technical support. We tested different machine learning models and ways to handle the imbalance in our dataset. The best results came from a logistic regression model using class weighting, which was able to identify most churners while keeping a balanced overall performance. This makes it the most useful model for predicting and allowing a company to take action before a customer decides to leave.

That could mean targeted retention efforts, such as special offers or incentives, and predictions from this model would help select the customers that would be the most beneficial to target.

The model should be monitored over time to make sure it stays accurate in case customer behavior changes over time. Also, collecting more detailed customer data in the future, such as customer feedback or history of customer support could help improve the model even further. In the long-term, it could be worth trying more advanced models like boosted trees or neural networks that may lead to even better performance. But as for now, this logistic regression model provides a solid foundation for reducing churn and supporting customer loyalty.