Telco Customer Churn

A Data Analysis Project

Term Project - Group 2 SCS 3251

April 16, 2019

Kerim Terzioglu (kterzioglu@yahoo.com)
Jitendrakumar Prajapati (jeetprajapati@gmail.com)
Masthanaiah Pelluri (mast311@gmail.com)
Jinsong Shi (shijinss@gmail.com)

Preface

The following analysis work forms the term project, submitted as a combined *group* effort by the aforementioned members of University of Toronto - School of Continuing Studies. This *group* has undertaken the project as a mandatory requirement for the course - Statistics for Data Science, Section 015 in the Winter 2019 semester.

In order to avoid any possible copyright infringements, the analysis has been done on a data set publicly available on Kaggle, named - Telco: Customer Churn. The data set was approved by the instructor Sergiy Nokhrin to be used for the given purpose in the current context.

Any questions or concerns regarding the project can be directed to any of the group members listed on the previous page.

Objective

In this project we have used be the Telco dataset to study customer behaviours in order to develop focused customer retention programs.

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. They need to understand who is leaving. The objective of our study is to find out who is leaving and why.

This project builds a predictive model which was covered in the course. The analysis was structured in the following fashion:

- Data Exploration / Feature Engineering
- Graphical Analysis
- Machine Learning Models
- Results

Some of the questions that we will try to answer during this project are:

- Which variables influences client departures?
- What are the most important variables to consider?
- Which attributes have the highest probability for customers who churn?

The objective of this study is to look at the available data set and to attempt to answer some of these questions for the benefit of Management who is looking for insights for refining their customer retention programs. So let's proceed further and look at what this report has to offer.

Data Preparation

The data set used for the analysis is the Kaggle Telco dataset. Overall the data quality was excellent (only 11 missing values). Python was used for data exploration and regression analysis.

This sample dataset contains information about Telco customers and if they left the company within the last month (churn). Each row represents a unique costumer, while the columns contains information about customer's services, account and demographic data.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

This dataset has 7043 samples and 21 features.

There are only 11 missing values, all of them for the TotalCharges column. Given these values are actually blank, we assumed zero charges were associated to these customers as they were with no previous tenure, hence we substituted the missing values with zero values.

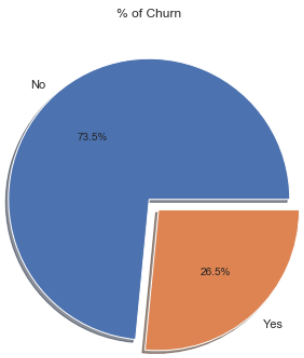
Additional steps were taken to remove the column customerID (all occurrences were unique in values which were deemed to be not useful for predication purposes).

Churn was converted to binary values (0 - no churn, 1 - churn)

Categorical columns were converted to dummy columns.

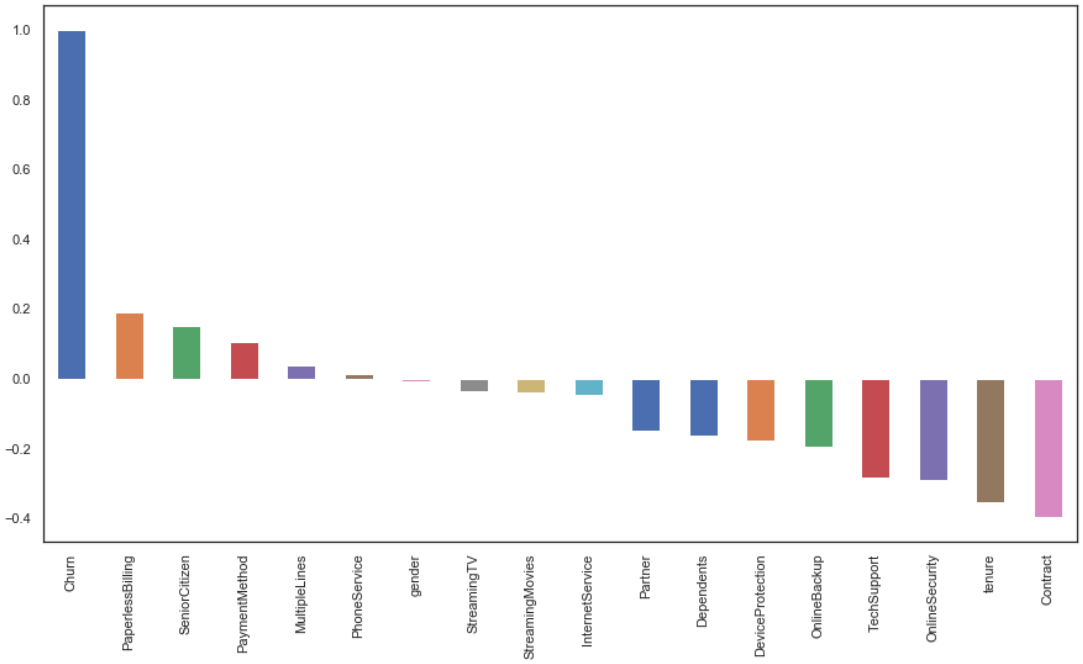
Analysis

On a periphery, we observe that 27% customers have left the Telco service. Which is a very high percentage of churn.

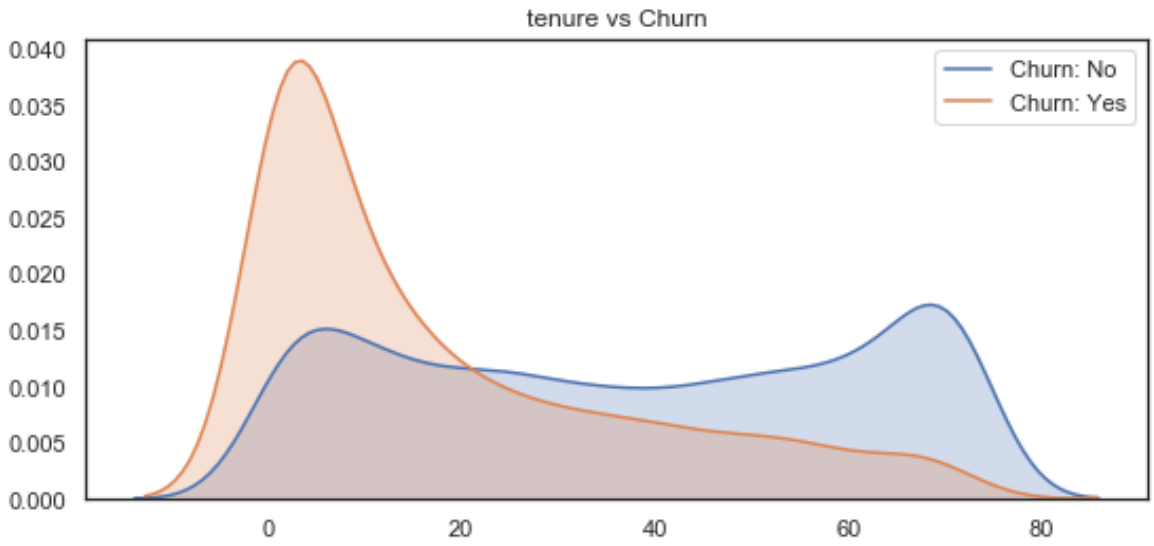


We observe a correlation between variables:

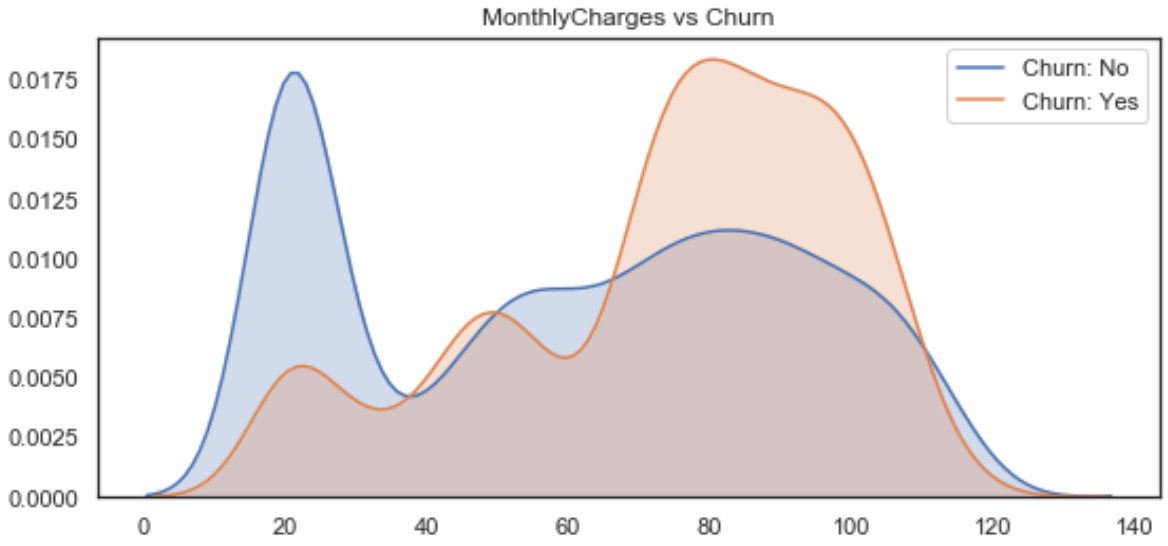
- Month to month contracts, absence of online security and tech support, Fiber optic Internet services are having positive correlation with churn.
- Surprisingly the paperless billing is having positive correlation with churn.
- Tenure, Two year contracts are having negative correlation with churn.
- Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection have negative impact to churn.



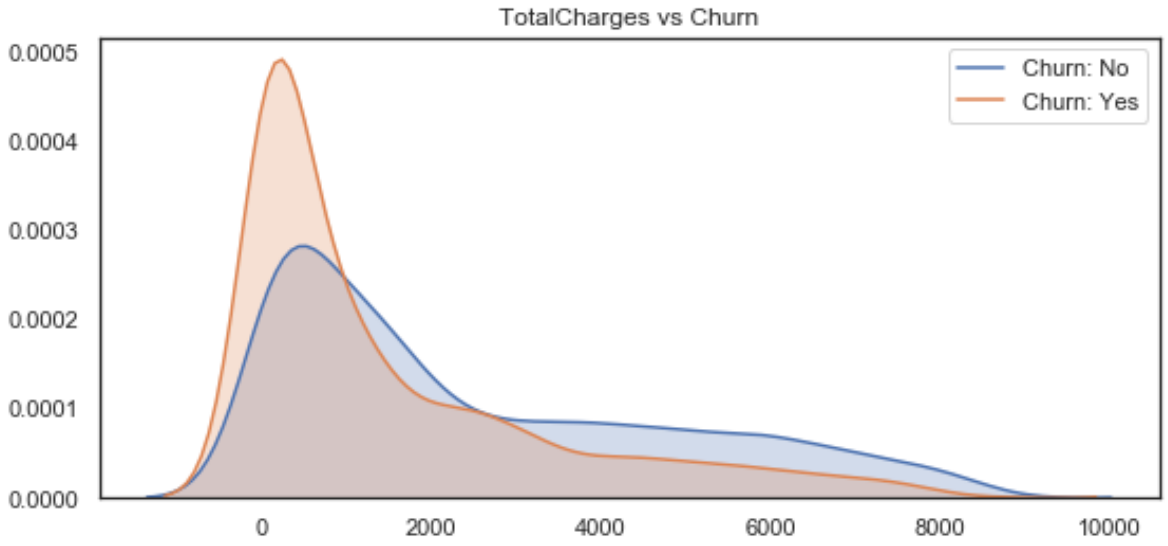
Customers who have recently joined the Telco Network have very high chances of leaving the network:



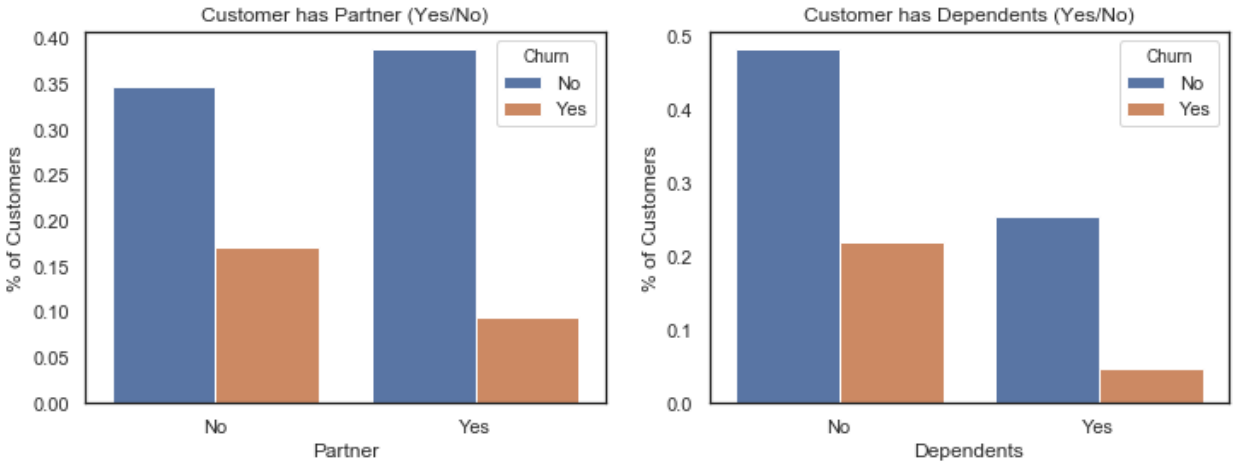
Customers who are paying high monthly charges are also likely to leave the network:



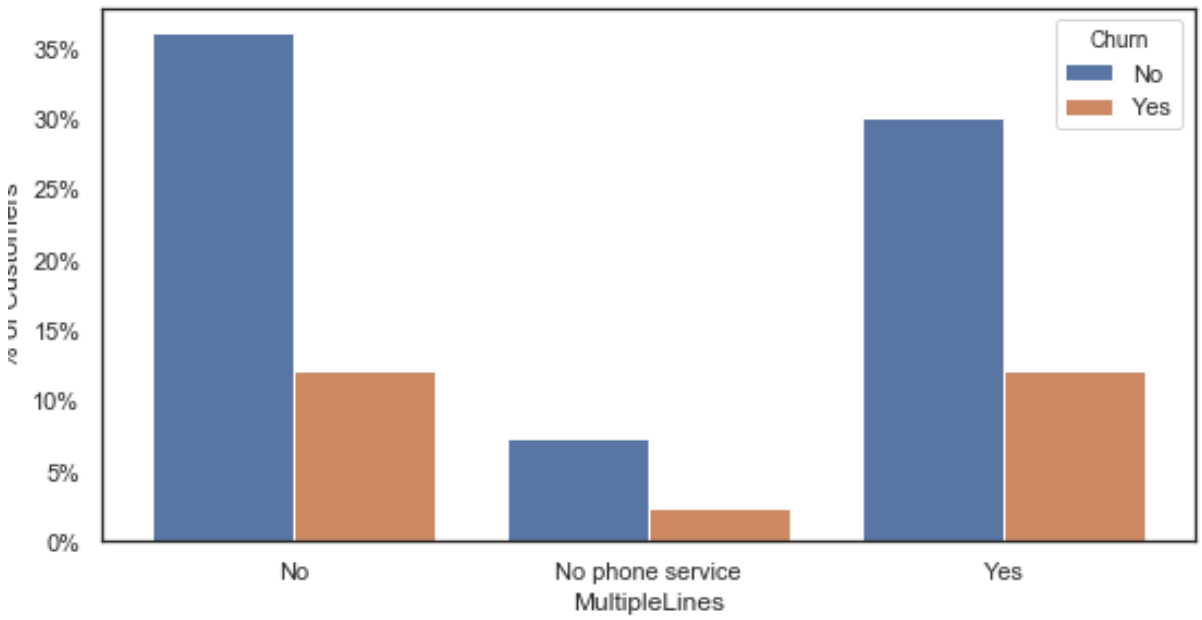
Lower Total Charges also result in high churn rates:



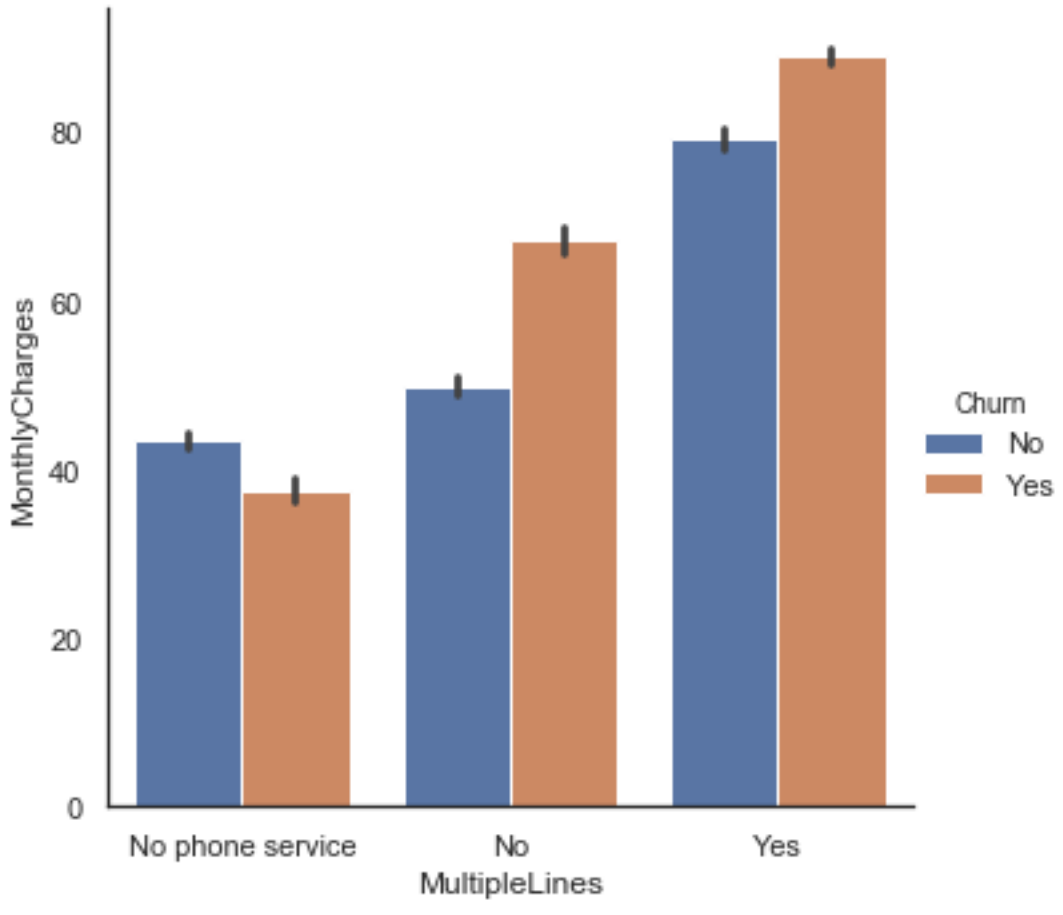
Customers who don't have parters or dependents are more likely to churn:



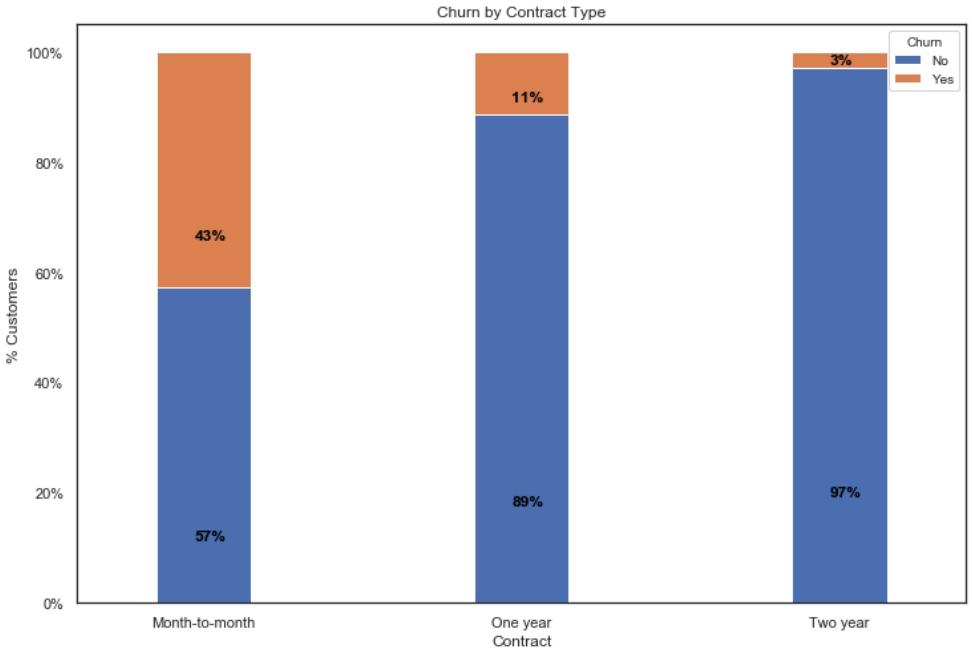
Customers with multiple lines have higher rates of churn:



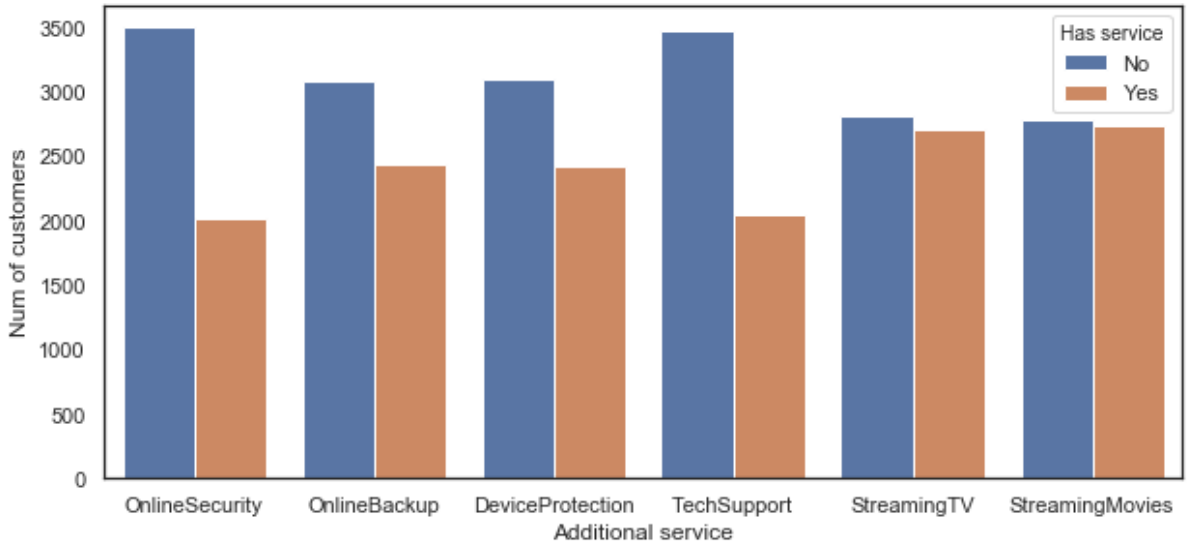
Customer with multiple lines and higher monthly charges are likely to leave the network:
:



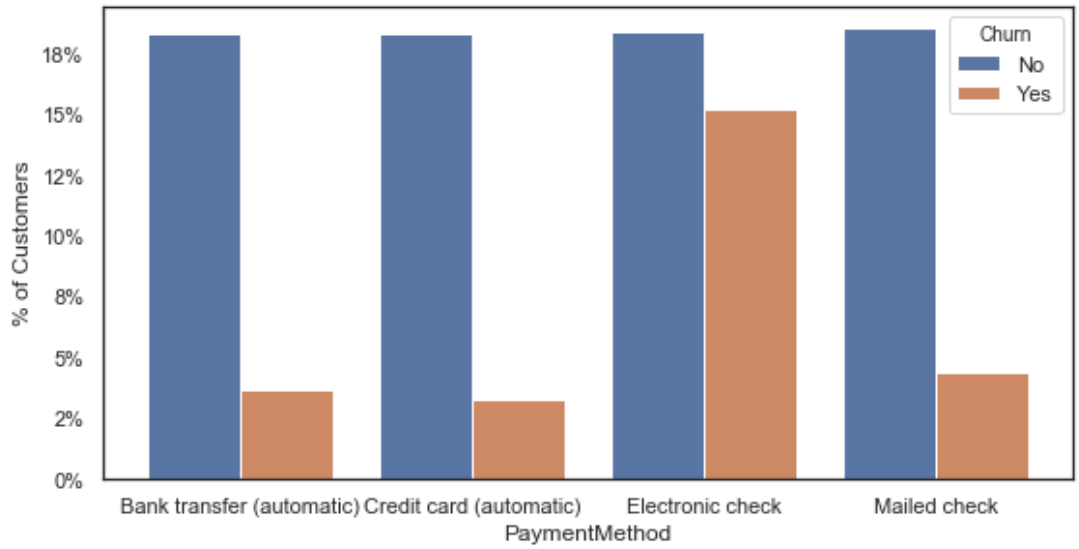
Customer with shorter terms contracts (month to month) demonstrate very high rates of churn:



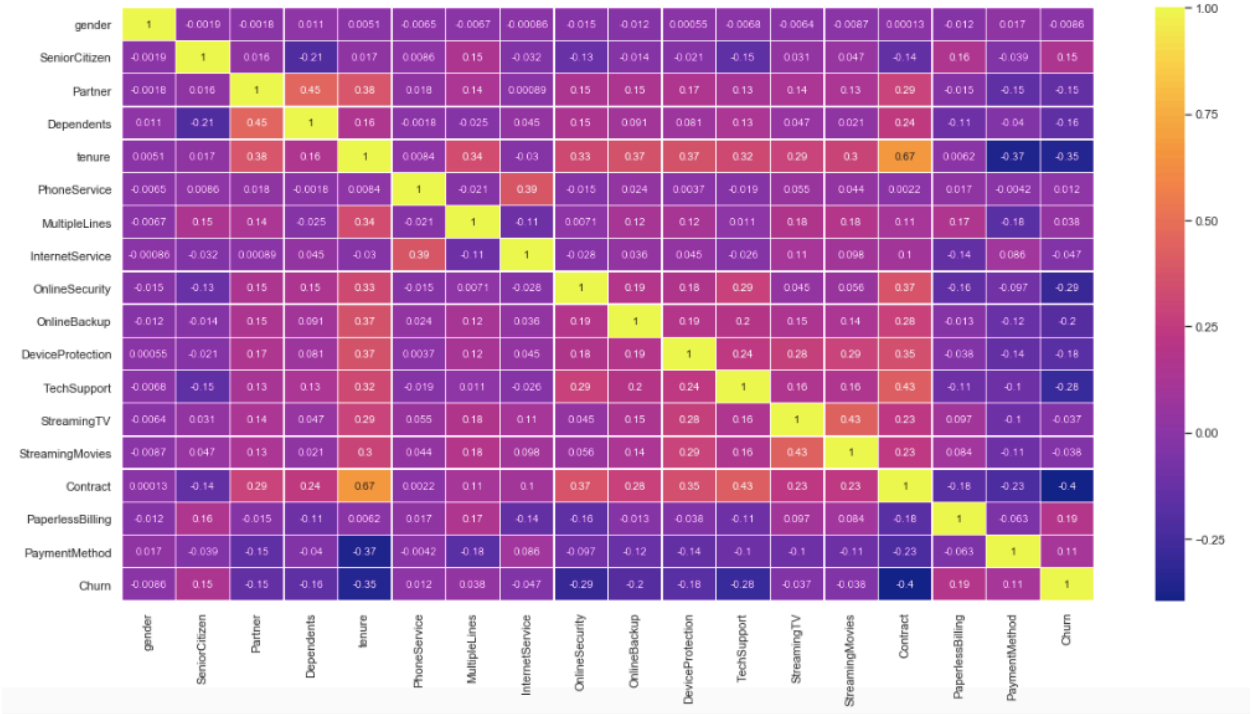
Customer who have streaming services are more likely churn:



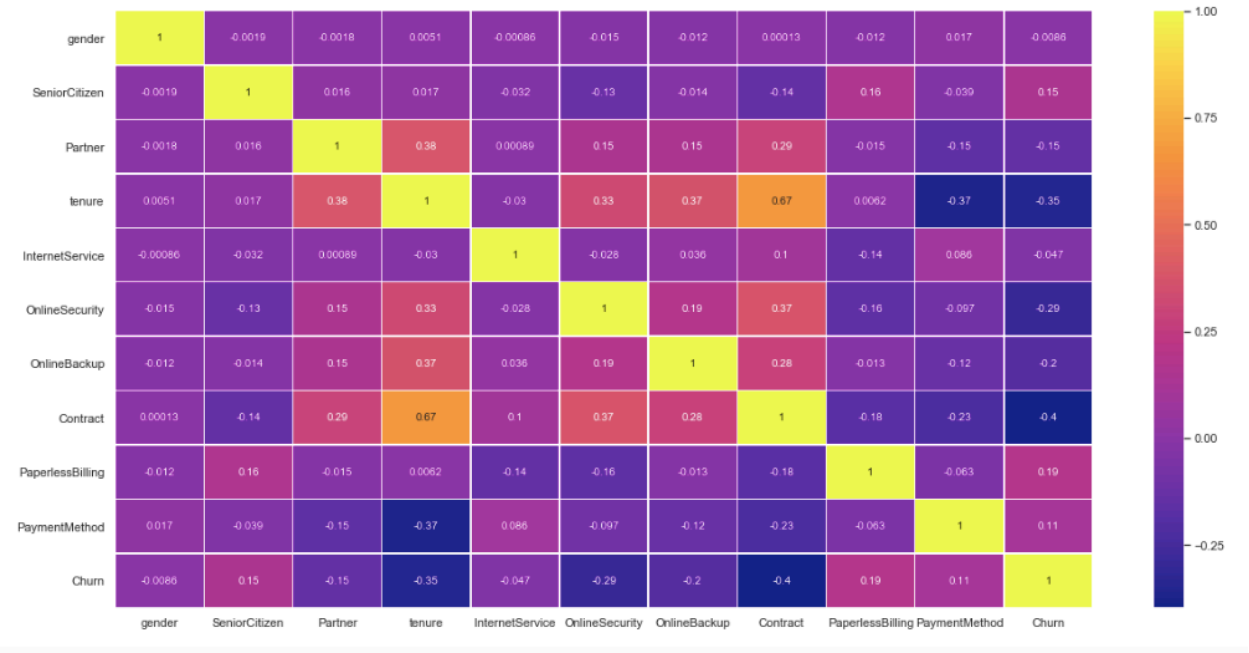
Customers using online payment methods are more likely to churn:



A preliminary correlation "heat-map" was built to determine the factors with maximum influence on churn This gave us a clear indication as to which parameters were to be kept at the core of the analysis.



With the above heatmap, we further restricted the number of attributes to the top 10.



Our analysis then used two classifiers, Decision Tree and Logistic Regression, using the top 10 attributes selected above.

Predictive Modelling

For this exercise, we evaluate the performance of 2 machine learning algorithms by using different training and testing datasets.

Logistic Regression Classifier

We took our original dataset and split it into two parts. We first trained the algorithm on the first part, made predictions on the second part and evaluated the predictions against the expected results.

The size of the split used was 80% of the data for training and the remaining 20% for testing. We then tested and evaluated the accuracy of a Logistic Regression model.

In working with the model parameters, we choose:

- max_iter of 100,000
- Utilized different C value of 1000,100,10,1,0.2, 0.1, 0.05, 0.02, 0.01 to train the model until obtaining the best accuracy.

In addition to specifying the size of the split, we also specify the random seed. Because the split of the data is random, we want to ensure that the results are reproducible.

Model Evaluation:

We observe that the estimated accuracy for the model was approximately 79% when a C of 0.02 is used:

```
Accuracy(%):  0.7877927608232789
Mean Squared Error:  0.21220723917672107
Explained Variance Score:  -0.08050608641604606
R2:  -0.10954232047574641
```

Here, **variance** is a measure of how far observed values differ from the average of predicted values, i.e., their difference from the predicted value mean. The goal is to have a value that is low.

R2 describes the amount of variation in the response that is explained by the least squares line or model. It is used to describe how well the model fits the data.

mean square error (MSE)—is the average of the square of the errors. The larger the number the larger the error. **Error** is this case means the difference between the observed values and the predicted ones.

Decision Tree Classifier

Similar to above, we split our data. The size of the split used was 80% of the data for training and the remaining 20% for testing. We then tested and evaluated the accuracy of a Decision Tree model.

In working with the model parameters, we choose different max-depths to train the model:

- set hyperparameter max_depth of 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16, 20, 30, 40, 50, 60, 70, 80, 90, 100 and None

Model Evaluation:

We obtain that the accuracy is the best when max-depth is 5

	Depth	Accuracy	Mean Squared Error	Explained Variance Score	R2
2	2	0.729595	0.270405	-0.343857	-0.413832
3	3	0.770759	0.229241	-0.191205	-0.198603
4	4	0.777147	0.222853	-0.0960869	-0.165205
5	5	0.779276	0.220724	-0.132263	-0.154072
6	6	0.775018	0.224982	-0.173807	-0.176338
7	7	0.773598	0.226402	-0.175792	-0.183759

Next, we utilized hyperparameter to obtain the best max-depth which resulted in the following:

```
LogisticRegression {'max_depth': 5}
Accuracy:  0.7792760823278921
Mean Squared Error:  0.2207239176721079
Explained Variance Score:  -0.13226300902296018
R2:  -0.15407244705002388
```

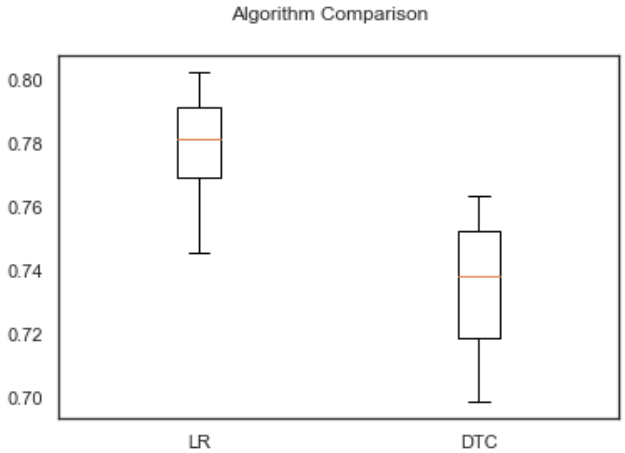
Given both models had similar scores for accuracy, MSE, Variance and R2, we turned to the SciKit helper function to validate our findings.

model_selection.cross_val_score was used to evaluate both models using a K-fold cross validation. Cross validation is an approach that is used to estimate the performance of a machine learning algorithms with less variance than a single train-test set split.

It works by splitting the dataset into k-parts (k=10). Each split of the data is called a fold. The algorithm is trained on k-1 folds with one held back and tested on the held back fold. This is repeated so that each fold of the dataset is given a chance to be the held back test set.

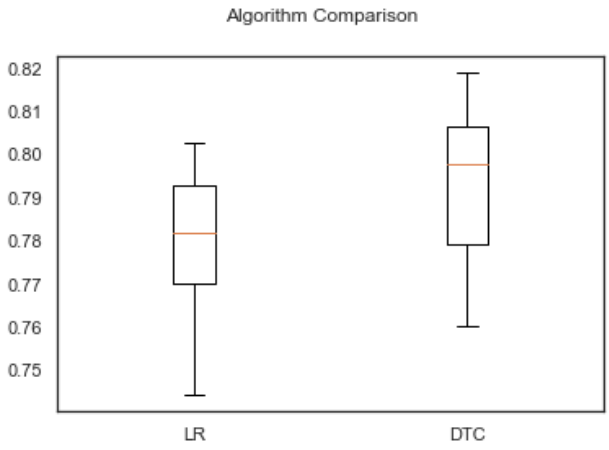
When running the cross validation utility, we end up with k different performance scores that can be summarized using a mean and a standard deviation. The result is a more reliable estimate. It is more accurate because the algorithm is trained and evaluated multiple times on different data.

For our exercise, we ran the helper function twice. First, we used models that were not fitted and observe the LR algorithm performed better.



The second run run utilized models that were fitted.

Here we observe, the Decision Tree classifier is the better algorithm:



Model Prediction

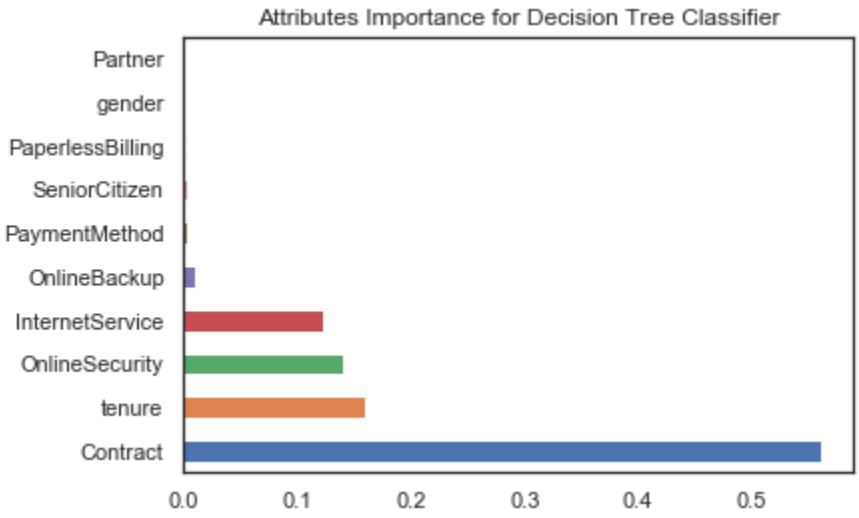
We proceed to make predictions on our validation dataset with the Decision Tree classifier yielding the following results:

	precision	recall	f1-score	support
0	0.80	0.81	0.80	1046
1	0.44	0.43	0.43	363
micro avg	0.71	0.71	0.71	1409
macro avg	0.62	0.62	0.62	1409
weighted avg	0.71	0.71	0.71	1409

This above table indicates our model has a good overall prediction.

However, what is more important is not to predict exactly the clients that will leave but to have the probability of the client leaving or not.

We can still get some a information from this model about the features that are most important, for that we examine the coefficients of the model:



The above plot gives us the most important attributes to predict if the client will leave or not, where the highest ones indicate strong chance of leaving and the lower ones indicates high chance of staying.

Conclusion

With the numerical and categorical plots we were able to see easily which variables influenced client departures.

The regression model confirmed the most important features used to predict if the client will leave along with probabilities for each of the associated attributes.

Based on the data and predictions, this Telco may potentially experience a loss of 1,869 clients with a total revenue loss of $2,736,826.