Predicting Churn

Predicting customer churn

By Steven Jones

13/08/20

Introduction

The aim of this project is to use machine learning to predict which customers are most likely to leave us soon.

The process to do this will involve analysing our data in detail to see what the drivers are. As well as guiding us in understanding which machine learning models would be best placed to identify churn, it will also help us understand our business, the areas we need to focus on and actions we need to take to help us reduce churn.

The dataset used is a publicly available set from Kaggle.com. I have already split this data into a training set (90%) and a test set (10%). I will use the train set to create the machine learning model and then judge its accuracy by predicting customer churn in the test set.

Project goals

We have 3 measures of success to look out for:

Accuracy is the total accuracy for the model which means how many customers who churned / didn't churn did we correctly identify.

Sensitivity is the 'true-positive'. It is the accuracy in which we correctly identified just the customers who churned.

Specificity is the 'true-negative'. It is the accuracy of identifying customer who didn't churn.

We will delve into, and understand why, each of these are important as we go through.

Exploratory Data Analysis

The data has 7,032 observations from 21 variables. The variable columns are:

- Customer ID

Demographic data

- Gender
- Senior Citizen
- Partner
- Dependents

Services

- Phone
- Multiple lines
- Internet service
- Online security
- Online backup
- Device protection
- Tech support
- Streaming TV
- Streaming movies
- Paperless billing

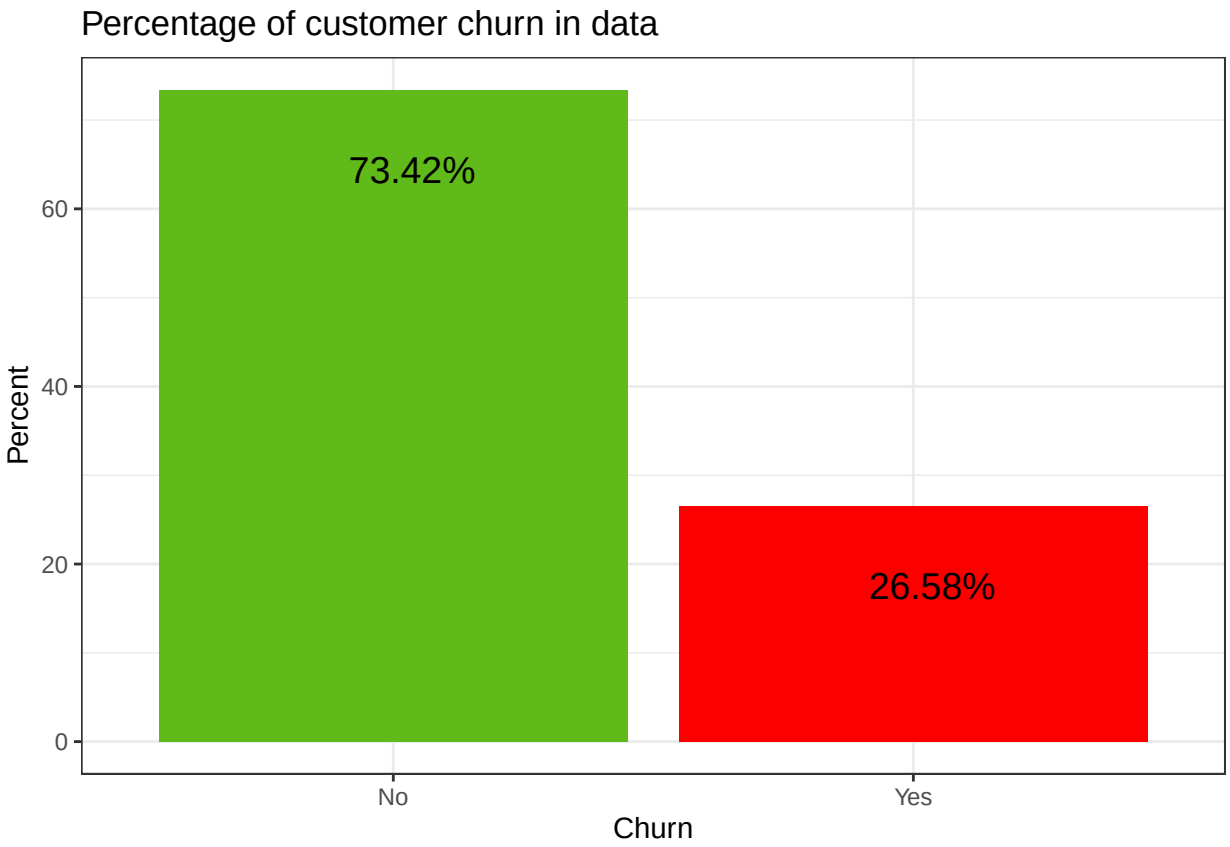
Contract / payment information

- Contract duration
- Contract type
- Payment method
- Monthly cost
- Total accrued cost

Churn

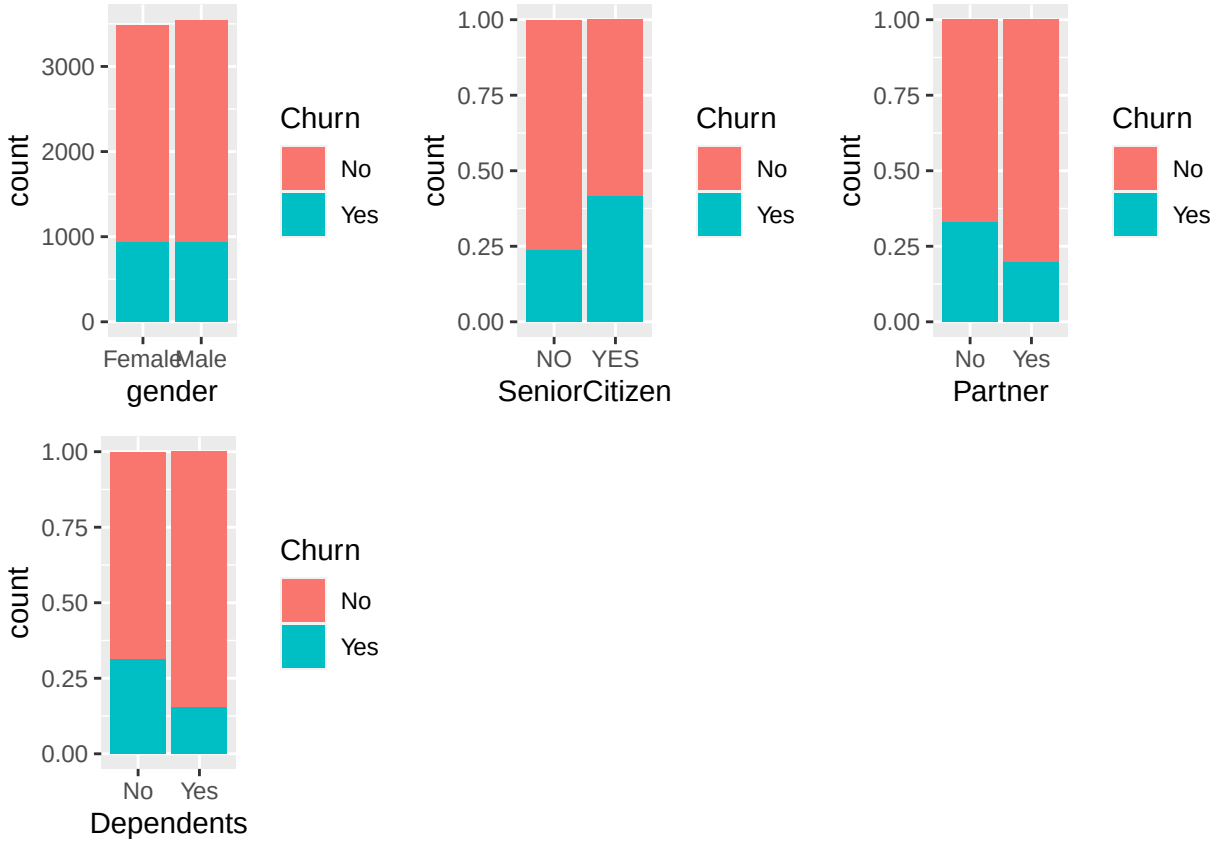
- Churn

Around 26% of customer left the platform:



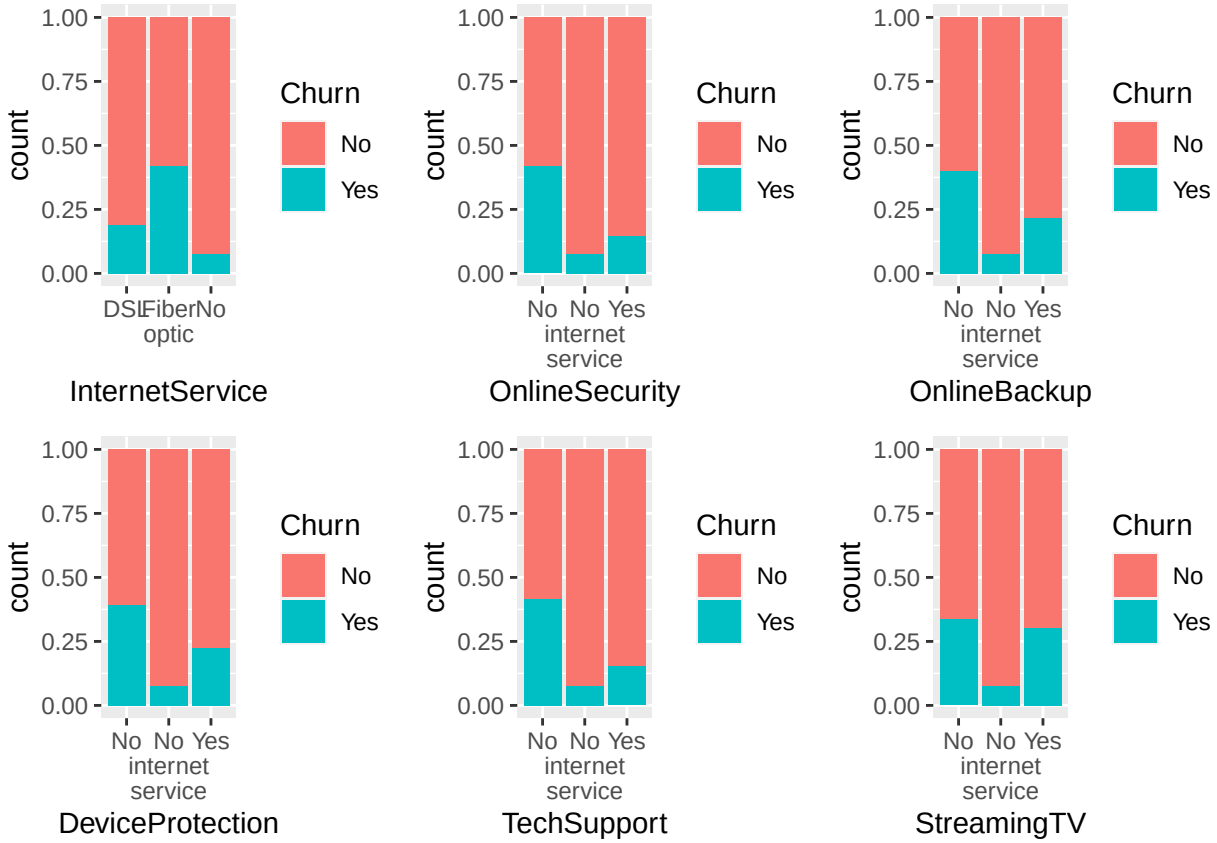
Demographic data

- Senior citizens were significantly more likely to churn
- Customers without a partner are more likely to leave
- Customers without children are more likely to leave



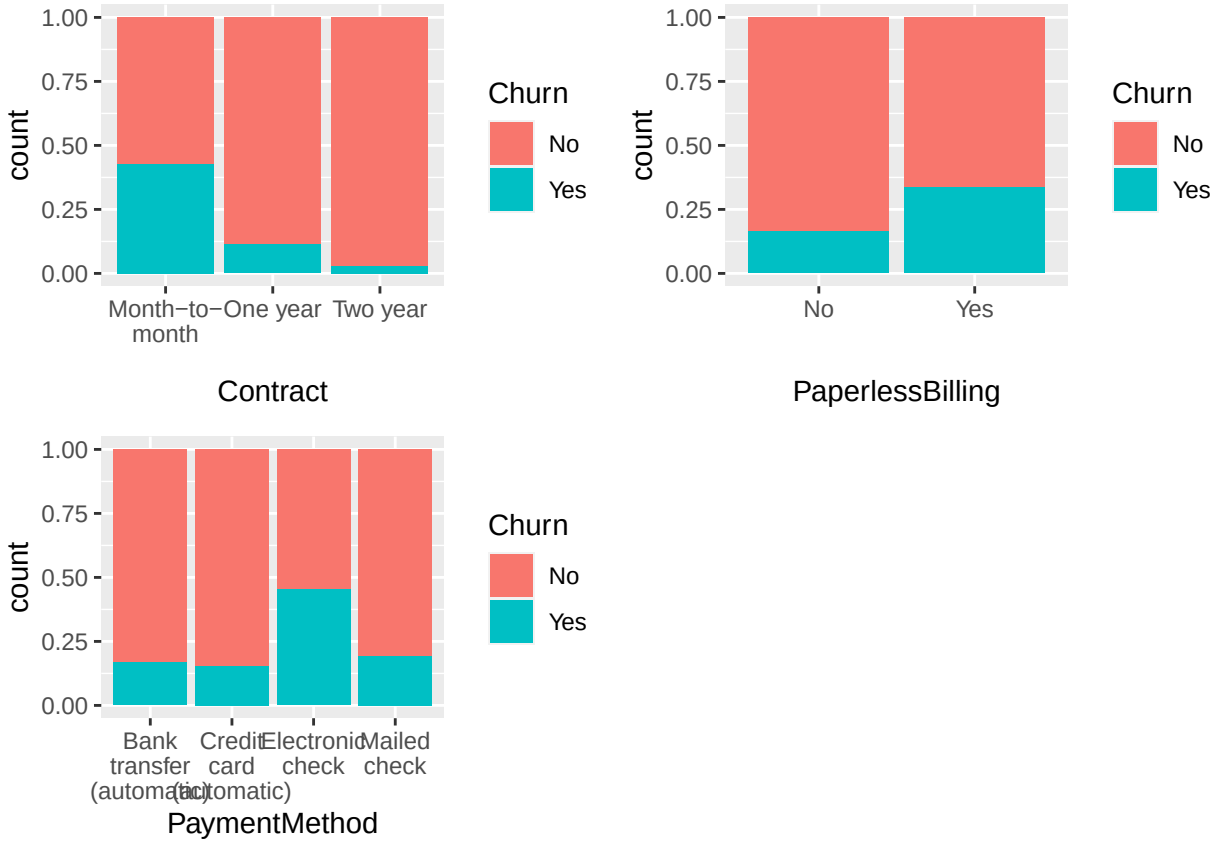
Services data

- Fibre optic customers are more likely to leave
- Customers without Online security, Online backup, Tech support or device protection are more likely to leave



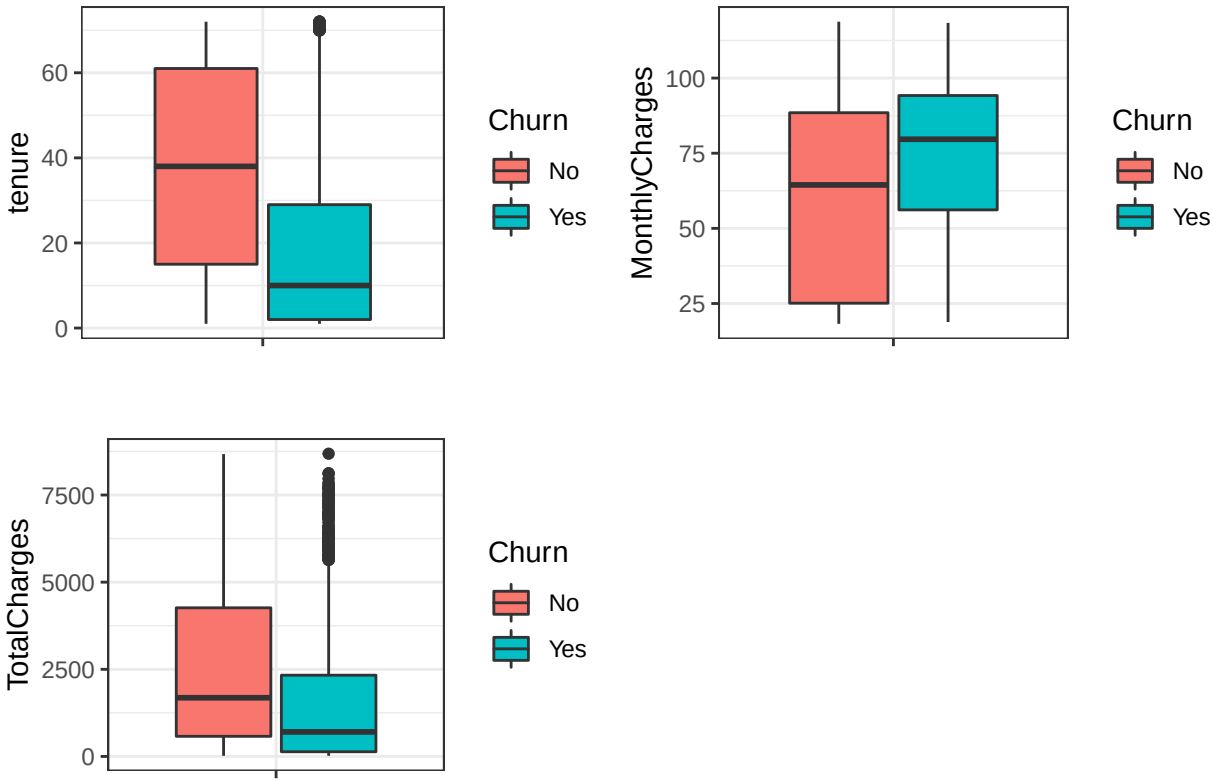
Contract / payment information

- Customers on a month to month contract are far more likely to leave
- Paperless billing customers are more inclined to leave
- Customers who pay via electronic check are most likely to leave



We also have 3 continuous variables (Contract duration, Monthly and total charges) which we can analyse.

- Customers tend to leave within the first 30 months
- The average (Median) point people leave is 10 months
- The average price point of customers who churn is about 81. About 20 higher than the average of customers who haven't churned



Building machine learning models

Now we know more about the drivers which cause customers to leave we can apply these to create a model which predicts if a customer is likely to leave.

The first model we will use will be a linear regression model in which each customer will be assigned a probability of leaving based on the criteria we've just looked at. We start with a 50/50 probability for each customer and then apply a + or - a certain number depending on each variable. A senior citizen might get a + 3% for example because we know they're more likely to leave than non senior citizens. We know after 30 months the customer is more likely to stay so they may get a -10%. We add these numbers up for each variable and if the final figure is above 50% we predict leave, under we predict stay.

The results of this are a 79% accuracy and I've correctly identified 55% of the customers who will leave within the next month.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1364  249
##        Yes  185  312
##
##                Accuracy : 0.7943
##                  95% CI : (0.7764, 0.8114)
##     No Information Rate : 0.7341
```

```
##     P-Value [Acc > NIR] : 7.788e-11
##
##                   Kappa : 0.4532
##
##  Mcnemar's Test P-Value : 0.002494
##
##             Sensitivity : 0.5561
##             Specificity : 0.8806
##          Pos Pred Value : 0.6278
##          Neg Pred Value : 0.8456
##              Prevalence : 0.2659
##          Detection Rate : 0.1479
##    Detection Prevalence : 0.2355
##       Balanced Accuracy : 0.7184
##
##        'Positive' Class : Yes
##
```

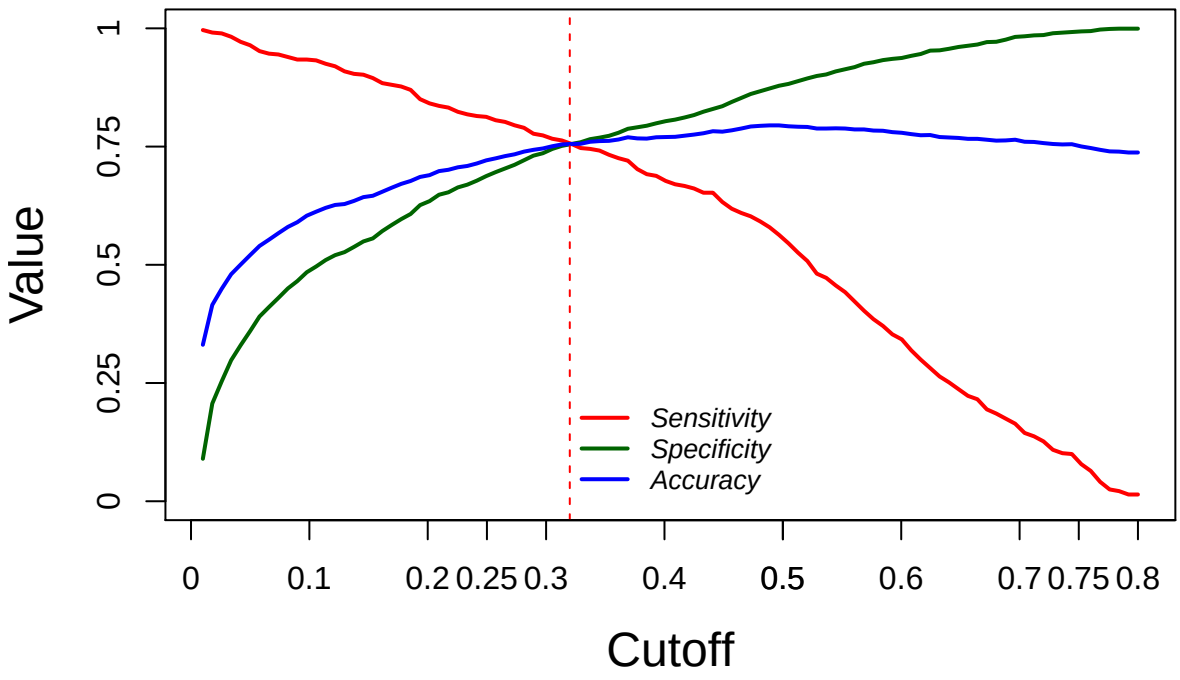
We can also look at the variance inflation factor. Numbers close to 1 indicate there is no correlation (between this and the chances the customer leaves in the next month) with higher numbers indicating more correlation.

We can see tenure has a high correlation and if the customer uses the fiber optic internet connection and thus these are the 2 areas to focus on if you were looking to decrease churn.

```
##                          tenure                  MonthlyCharges
##                        2.282029                        1.367104
##                    SeniorCitizen                         Partner
##                        1.075498                        1.085078
##    InternetService.xFiber.optic             InternetService.xNo
##                        1.519656                        1.509683
##                  OnlineSecurity                    OnlineBackup
##                        1.087319                        1.150754
##                     TechSupport                     StreamingTV
##                        1.145909                        1.291089
##              Contract.xOne.year              Contract.xTwo.year
##                        1.203896                        1.212871
##                PaperlessBilling PaymentMethod.xElectronic.check
##                        1.111438                        1.128411
##           tenure_bin.x1.2.years           tenure_bin.x5.6.years
##                        1.051896                        2.005468
```

The final process for this function is to optimise the figures and even out the accuracy/sensitivity/specificity measurements. We can do this by finding the optimal probability cuttoff (this is how high does the probability need for us to predict a customer will leave). Lets check this on a graph:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006041 0.042040 0.194187 0.270761 0.486755 0.813298
```



The best cuttoff is around 0.32. Lets run the model with this and see what our final figure will be:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1170  136
##        Yes  379  425
##
##               Accuracy : 0.7559
##                 95% CI : (0.737, 0.7741)
##    No Information Rate : 0.7341
##    P-Value [Acc > NIR] : 0.01196
##
##                  Kappa : 0.4506
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.7576
##             Specificity : 0.7553
##          Pos Pred Value : 0.5286
##          Neg Pred Value : 0.8959
##              Prevalence : 0.2659
##          Detection Rate : 0.2014
##    Detection Prevalence : 0.3810
##       Balanced Accuracy : 0.7565
```

```
##
##        'Positive' Class : Yes
##
```

We can see accuracy, sensitivity and specificity are all at 75%.

Random Forest

Most people have come across a decision tree. In this case it might look like:

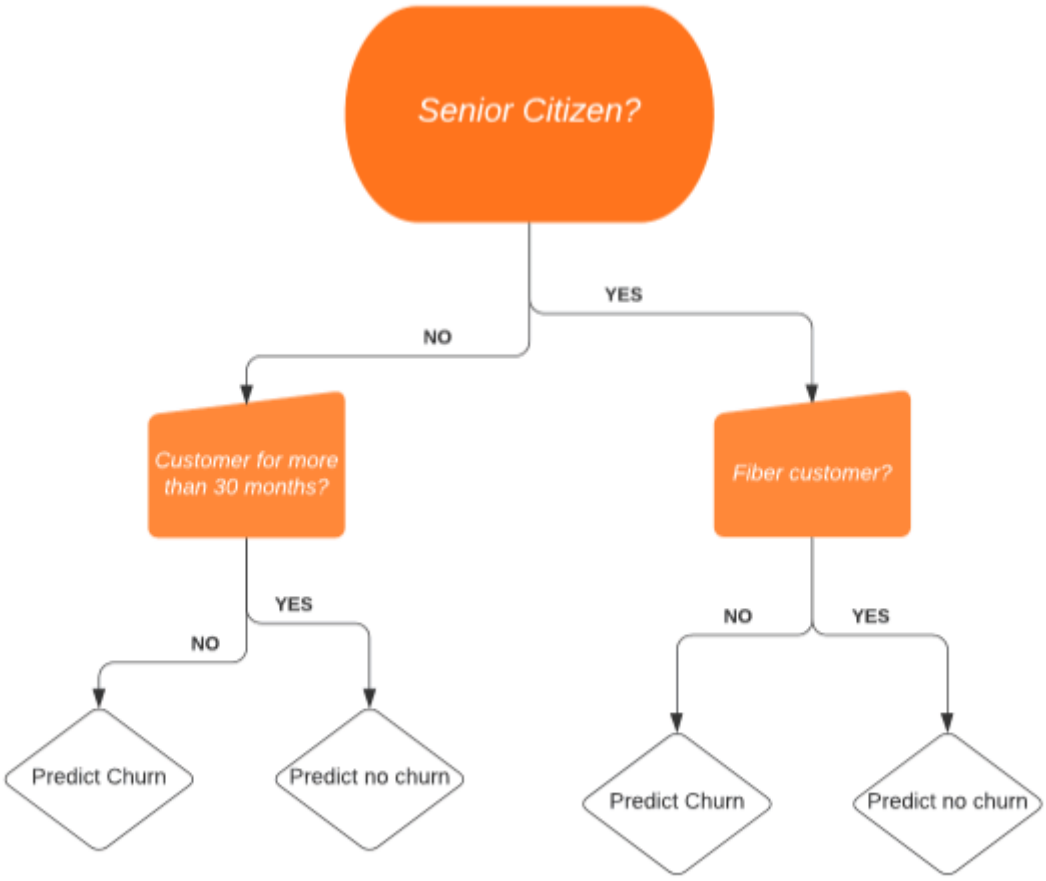


Figure 1: Example Decision Tree

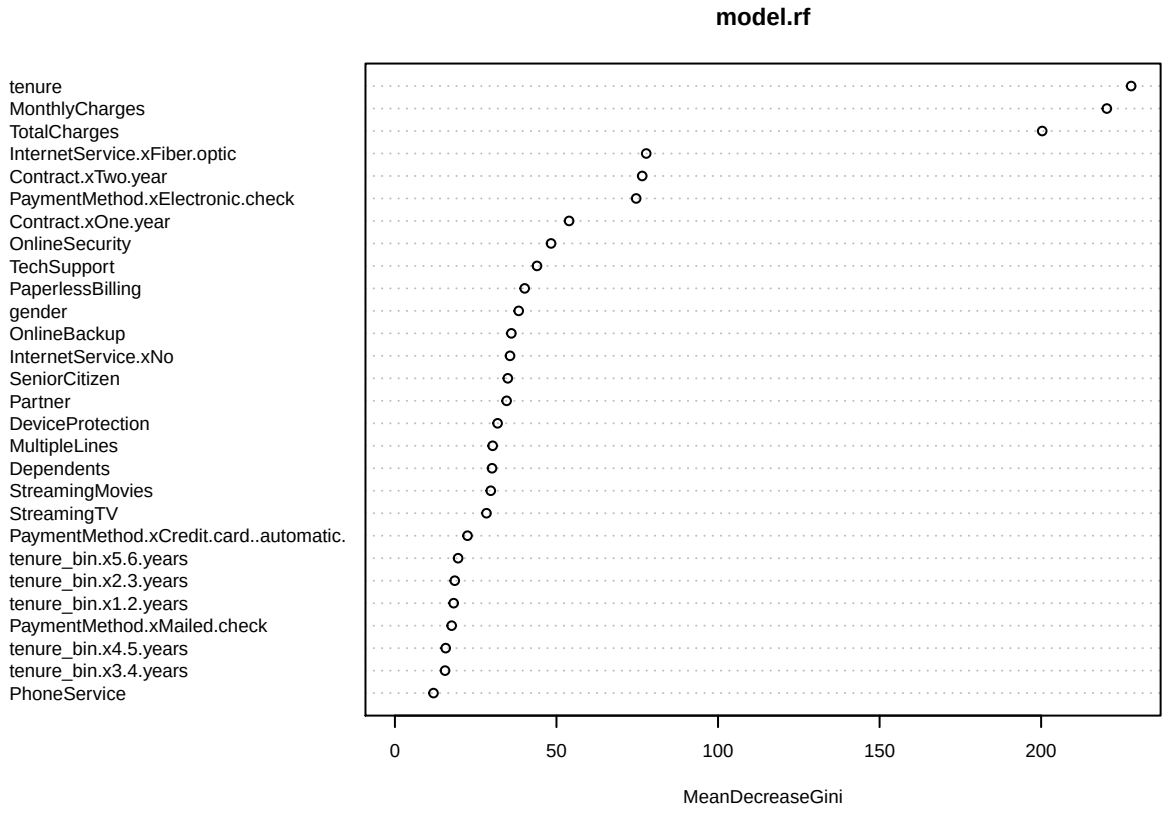
Random forest takes different samples from the data and runs decision trees from these samples, aggregating the result to determine the prediction.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1404  145
##          1  301  260
##
##                Accuracy : 0.7886
```

```
##                  95% CI : (0.7706, 0.8059)
##     No Information Rate : 0.8081
##     P-Value [Acc > NIR] : 0.9884
##
##                   Kappa : 0.4058
##
##  Mcnemar's Test P-Value : 2.145e-13
##
##             Sensitivity : 0.8235
##             Specificity : 0.6420
##          Pos Pred Value : 0.9064
##          Neg Pred Value : 0.4635
##              Prevalence : 0.8081
##          Detection Rate : 0.6654
##    Detection Prevalence : 0.7341
##       Balanced Accuracy : 0.7327
##
##        'Positive' Class : 0
##
```

The random forest improves our accuracy to 78% and, more importantly improves Sensitivity to 82%.

A quick look at the variable importance chart shows how monthly charge is also an important driver of churn.



Conclusion

Through applying Data Science to evaluate churn I've managed to create a model which will identify over 82% of the customers who will churn in the next month.

In the exploratory data analysis, I also identified the major driver behind the churn which gives us the actionable intelligence to make changes to our business and reduce churn.

The accuracy could be improved further through considering other machine learning algorithms and possibly through an ensemble of them. If I was to continue, I would also look to optimise the Random Forest method.