Predict Telco Customer Churn

Yuman Xie

May 2022

Special thanks to: Professor Changmi Jung, Claudia Dou

Business Understanding

We want to decipher how to reduce customer churn rate for companies. To most businesses, number of the customers is one of the most important keys to higher revenue and business success. A high customer churn rate means many existing customers are leaving the company, which implies the company is not satisfying customers' needs, and is earning less revenue. We are curious about what factors might lead to customer churn, and we would like to leverage data science knowledge to help companies reduce customer churn rate.

Customer churn occurs when customers or subscribers stop to receive service or subscription from a company. It directly affects the a company's profitability, and is a critical measure because it's less expensive to retain existing customers than to acquire new customers.

There are multiple reasons behind customer churn. Our project focuses on analyzing existing customers' data to find out potential factors lead to their churn, and predicting which kind of customers are more likely to churn. Such understanding allows us to identify customers with high churning possibility, and provide companies with targets and time to retain those customers.

To achieve such goal, we need a set of data that contains customers' business-related features, and whether the customers left the company or not, as targeting variable. According to this standard, we used the dataset from a telephone company Telco: Telco Customer Churn. In the competitive telecommunication industry, reducing customer churn is critical for higher revenue.

In this report, we will discuss how we prepared, explored, visualized the dataset, to identify possible important customer characteristics (features in the dataset) that might affect churn rate. We will include the variable importance analysis through EDA, to help Telco identify important reasons of customer churn; Telco can potentially alleviate these reasons, and potentially extend business to customers without these concerns. Then we will talk about our models predicting customer churn, and model validation results. From there we will explain how our results, as high as 80.9% accuracy and 53% recall, are useful for Telco, and how Telco can use our models to predict customers likely to churn, and prepare strategies to retain them. Finally, we will elaborate the major concerns Telco might have regarding our models, as well as our plans of mitigations.

Data understanding & visualization

We used the dataset Telco Customer Churn. Exhibit 1 indicates the source of our data. The dataset has 21 attributes and 7043 instances. Exhibit 2 explains, in detail, all the 21 attributes. Because our focus is to predict customer churn, we set the binary attribute "Churn" as our response variable, and the rest of the 20 variables to be explanatory variable. Among the 20 explanatory variables, there are 4 numeric variables, and 16 categorical variables. The dataset is neither balanced nor heavily imbalanced, with 73% no churn and 27% churn.

Data Processing

Before utilizing it, we completed the following steps to prepare the dataset into the format required for our usage. First, we formatted variable names. Second, the variable "SeniorCitizen" is binary, but appears to be numerical in the original dataset, we factored this variable. We also figured there are 11 NA values in the column "totalCharges", because it's not considered a large amount of data points and won't affect our dataset too much, we chose to omit them directly. The variable "cusomterID" is a non-repeating index for the customers, it's of no value to us, thus we omitted it from the dataset. Variables "multipleLines", "internetService", "onlineSecurity", "onlineBackup", "deviceProtection", "techSupport", "streamingTV" and "streamingMovies" have values "Yes", "No", and "No phone service". For the convenience of further study, we converted "No phone service" to "No", and these variables are now binary variables with value "Yes" and "No".

Visualization

Exhibit 3 & 4 shows visualizations for Numeric Variables. From the summary in Exhibit 3, we can see that there are slightly more male than female (1.9%). Both female and male have average tenure of around 32 months, with male having 1 more month in average tenure (33). Female customers have slightly higher (1.3%) monthly charges than male. Both male and female have around $2283 average total charges.

Exhibit 4 is a correlation plot for numerical values. We can see that Total Charges is strongly correlated with Tenure, especially among customers who churn. Total Charges is also highly correlated with Monthly Charges. From the histograms we can see, tenure is relative uniform when churn is no, and left skewed when churn is yes. Distribution for Monthly Charges is variant, when churn is no it's more left skewed, when churn is yes it's more right skewed. For Total Charges, churn with both values yes and no have left skewed distribution.

Exhibit 5 to Exhibit 11 are visualizations for categorical variables. We can see that customers who churned have fewer months of tenure and higher average monthly charges (Exhibit 5). More customers pay by month-to-month using electronic checks, and these types of customers have the largest percentage of customers churned; customer under one-year or two-year contracts are much less likely to churn; customers using automatic payment methods (bank transfer and credit card), as well as mailed check are less likely to churn; customers using Fiber Optic are most likely to churn (Exhibit 6). Customer demography, specifically Senior Citizen, Partner, and Dependents, might affect customer churn; because Gender is very similar for customers both churn and not churn, we might choose to omit it in further modeling (Exhibit 7). Exhibit 10 is a combination of Exhibit 8 and Exhibit 9, so that we can compare the rest of the variables on same scale. We can see that Multiple Lines, Streaming TV, Streaming Movies have very similar patterns.

Online Security and TechSupport have very similar patterns. Online backup and Device protection have very similar patterns.

We also utilized Tableau to explore trends and potential factors that may influence customer churning. From Exhibit 12 we can see that a large percentage of customers are enrolled in low monthly Charges ($15-$28), and the churning rate is the highest among $68-$105. We also observed tenure's effect on customer churning. In Exhibit 14, as the tenure increase, customer churn (orange line) steadily decreases. This proves the negative relationship between churning and tenure. We also compare the size of churn of not churn based on payment method, which we can see an obvious large portion of people who use electronic check churned.

Variable selection

From the result of EDA, we decided to omit Total Charges, due to multicollinearity. We think Gender, Multiple Lines, Streaming Movies, Tech Support, and Device Protection might not contributing to model building either, but we chose not to omit them in the beginning.

Data finalizing

According to the result of variable selection, we omitted "totalCharges" from our dataset. We normalized the remaining two of our numerous variables ("tenure", "monthlyCharges"). The resultant values of "tenure" and "monthlyCharges" fall into the range of [0,1], so that the weights for these variables in our models may be interpreted as importance indicators. Before proceeding to modeling, we also split dataset into 70% of training dataset, and 30% of testing dataset, for model validation that'll be discussed in the later part.

Modeling

Our response variable is being labeled in the data as "churn", and is binary with value "Yes" and "No". Because of that, our problem is a supervised classification problem, thus we chose to compare the performance and select the best model from logistic regression, classification tree, random forest and SVM. All these methods are parametric classification techniques used to estimate the probability of an event happening, for instance, given relevant information and measure whether a customer will in fact leave Telco.

Logistic Regression

One of the biggest advantages of the logistic model is its interpretability of the model parameters. We were able to quantify the relationship between the response and our input features, according to the size of the coefficients and the significance of predictors. We can have a straightforward idea on which features might affect customer's choice more. However, Logistic Regression requires no multicollinearity between independent variables. In our dataset, many categorical variables may have unseen correlations, like the payment method and age, Senior Citizens. It is also hard to derive a complex relationship using logistic regression. Neural Networks, for example, would be a good alternative that could solve this drawback.

After running this model, we can see the coefficients, their standard errors, and the associated p-values. Tenure, Payment method, contract year one and two, internetService, streamingTV or movies, whether multiple Lines and seniorCitizens are statistically significant.

Variables with larger values of coefficient have more influence on response variable. For instance, for every one unit change in tenure, the log odds of churn (versus did not churn) decreases by -2.351. For one unit increase in monthly charges, the log odds of churn decrease by 8.193. The indicator variables about whether use internet service of fiber optic has a slightly different interpretation, it will increase the log odds of churn by 2.841 if use the fiber optic.

Classification Tree

We also used classification tree. The reason use classification tree is because it is intuitively easy to understand and present. This could help us better convey our model result to management level who have no prior knowledge on data. Classification tree splits the data into multiple sets and each set is further split into subsets to arrive at a tree like structure and make a decision. Homogeneity is the basic concept that helps to determine the attribute on which a split should be made. A split that results into the most homogenous subset is considered better and step by step each attribute is chosen that maximizes the homogeneity of each subset.

After we run our model, we can see from the tree (Exhibit 12) that only customers not choosing one year or two-year contract, having no of DSL as internet service, and have tenure larger than 0.18 will churn. We think customers who didn't sign contract will have less loyalty compared to customers who signed. We also think that the longer the customer has stayed with the company, the more likely they will keep Telco service. This is correspondent to our observation in Tableau visualization.

However, classification tree easily overfits, with possible high training but low testing performance. High variance is another problem causing poor model performance. To overcome these disadvantages, we decided to try Random Forest.

Random forest

Random forest is built using a combination of many decision trees, where each tree takes a random sample of the data with replacement and selects a random subset of predictors, resulting in a relatively uncorrelated set of decision trees. Each tree then makes a prediction and the class with the most votes becomes the model's final prediction. This is a great model for us since it would be more applicable for predicting the churning of new set of customers. We set model parameter to 500 trees, the number of variables randomly sampled as candidates at each split to 2.

One of the advantages of Random Forest is that, it gives out-of-bag (OOB) error estimates, which is the mean prediction error on training samples, using the trees that do not have that training sample in their bootstrap sample. It may act as a cross validation error and eliminate the need of using test/validation data, thereby increasing the training data.

SVM

In contrast to logistic regression, Support Vector Machines do not make any assumptions about the data. In SVM, we search for a hyperplane that separates the data as best as possible while maximizing the distance between the classes of our response variable.

Model Validation

For model validation, we used holdout validation, where 70% of the dataset is used for training and 30% dataset is used for testing. We used model accuracy and recall value (false negative rate). Accuracy predicts how many correct predictions were made from total number of decisions, which is a generic measure for model performance. Also, because our goal is to find out customers who might churn, it's best to have a higher recall value, which means we have less error in predicting customers as not churn, while in fact they will churn. It's less costly to pay more attention to those customers who will not churn, than to pay no attention to those customers who will churn.

Logistic Regression:

The accuracy for the logistic regression is 0.808. The recall for this model is 0.538, which means this model correctly predicts positive cases 53.8% of the time.

Classification Tree:

The accuracy for the classification tree is 0.787. The recall value for this method is 0.363, which means this model correctly predicts positive cases 36.3% of the time.

Random forest:

The accuracy of the random forest is 0.797, which is higher than that of classification tree. The recall value of random forest is 0.446, which means this model correctly predicts positive cases 36.3% of the time.

SVM:

The accuracy for the SVM is 0.809. The recall for this method is 0.542, which means the model correctly predicts positive cases 54.2% of the time.

Model tuning

Logistic regression - stepAIC

We used stepwise regression to tune logistic regression. First, we built an initial logistic model with all variables included. Then we will use stepwise feature selection methods with the function called 'stepAIC'. The function iterates through all the variables until the lowest AIC model among all models is discovered, we then used the variables in the lowest AIC model to build our best logistic regression. As a result, the method returns these variables: seniorCitizen, phoneService, multipleLines, internetService, onlineBackup, deviceProtection, streamingTV, streamingMovies, contract, paperlessBilling, paymentMethod, tenure, monthlyCharges. We validated the tuned logistic model, and got an accuracy of 0.807, and recall of 0.538. We can see that the accuracy and recall of the logistic model didn't change much. This might because the variables omitted were not significant in the original logistic model.

Random Forest - Hyper tuning & Mean decrease GINI

We can used the function tuneRF()in place of the randomForest() function to train a series of models with different mtry (number of variables randomly sampled as candidates at each split) values, we found that when mtry equals to 2, random forest has the lowest OOB Error (Exhibit 17). We also used Mean Decrease Gini to determine the importance of each variable in classifying the customer churn. The Mean Decrease Accuracy plot (Exhibit 18) expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented from descending importance. The mean decrease in Gini coefficient measures how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model. As a result, we can see that tenure, monthlyCharges and contract are the most important; paymentMethod and internetService are the second important, onlineSecurity, paperlessBilling and techSupport are relatively more important than the rest.

After tuning, Random Forest achieved an accuracy of 0.798, and recall of 0.497. We see there's 0.2% increase (from 0.7967 to 0.7981) in accuracy, and 5% increase (from 0.4462 to 0.4965) in recall.

Because SVM's performance was already the best, we didn't tune SVM.

Comparing four models' performances before and after tuning, we found that the SVM and tuned logistic regression have very similar accuracy (0.807 for LR and 0.809 for SVM) and recall value (0.538 for LR and 0.542 for SVM), thus we would recommend both models to Telco.

Deployment

In the previous part of this report, we have discussed the business problem we are trying to solve, the dataset and data science methods we used, as well as the performance of our methods. In the final part of our report, we would talk about the application of our methods, how our methods and results are helpful to Telco, and the limitations of and possible concerns towards our methods.

In the data exploring part, we figured that customers with fewer months of tenure, higher average monthly charges ($68-105), with shorter contracts, paying by electronic checks, using Fiber Optic, are more likely to churn. These results are proved by multiple methods (visualization, modeling), and thus are robust. Telco can use these variables to determine if an existing customer needs retention strategies, and if an incoming customer is of high value. They can also pay attention to these variables when record data for future using.

We developed four models, and two of them (SVM and tuned Logistic Regression) can provide around 80% of prediction accuracy, as well as around 54% of recall value. Telco can use these models to predict customer churn based on other characteristics. When Telco gets data for another period, they can feed the new data into our models, and deploy the prediction of our models to see which customers might churn. If the predicted churn rate is indeed high, Telco can then sort out the predicted churning customers, and develop marketing strategies to retain these customers.

While deploying our method, Telco should process their dataset according to our procedure. This includes omitting NA entries, formatting variable names, factoring categorical variables, standardizing variable values, and normalizing numeric variables.

However, there are some flaws in our methods. Our recall value is not too high (around 50%), which means among those customers who will churn, half of them will be marked as "not churn" by our models, and will be "ignored" by Telco. Also, because tenue is an important variable contributing to customer churn, when determining if an incoming customer is of high value, Telco should be aware that new customer's 0 tenure value will highly impact the customer's predicted churn. To this, Telco could potentially estimate an average tenure, using the customer's other characteristics, to mitigate the flaw.

Also, Telco should be careful on collecting and using the customer data. Data security is an important concern. Customers won't want their personal information, like partner, dependents, to be analyzed. Especially if Telco were to apply marketing strategies to specific customers.

Appendix

Exhibit 1

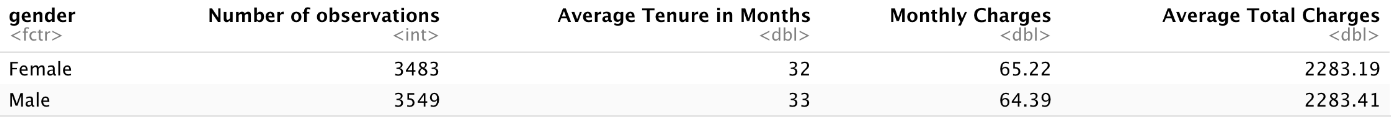
https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Exhibit 2

1. customerID: Customer ID
2. gender: Whether the customer is a male or a female
3. SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
4. Partner: Whether the customer has a partner or not (Yes, No)
5. Dependents: Whether the customer has dependents or not (Yes, No)
6. Tenure: Number of months the customer has stayed with the company
7. PhoneService: Whether the customer has a phone service or not (Yes, No)
8. MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
9. InternetService: Customer's internet service provider (DSL, Fiber optic, No)
10. OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
11. OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
12. DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
13. TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
14. StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)

15. StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
16. Contract: The contract term of the customer (Month-to-month, One year, Two year)
17. PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
18. PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
19. MonthlyCharges: The amount charged to the customer monthly
20. TotalCharges: The total amount charged to the customer
21. Churn: Whether the customer churned or not (Yes or No)

Exhibit 3



gender <fctr>	Number of observations <int>	Average Tenure in Months <dbl>	Monthly Charges <dbl>	Average Total Charges <dbl>
Female	3483	32	65.22	2283.19
Male	3549	33	64.39	2283.41

Exhibit 4

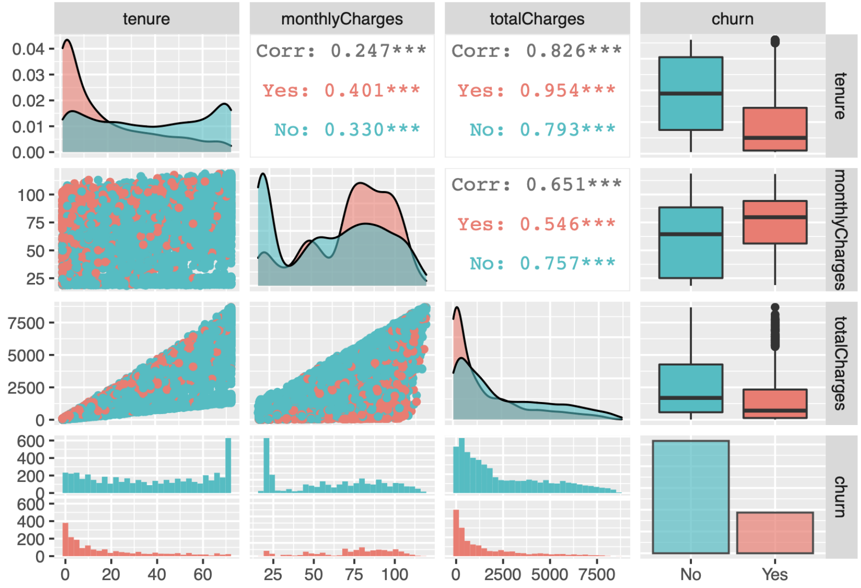


Exhibit 5

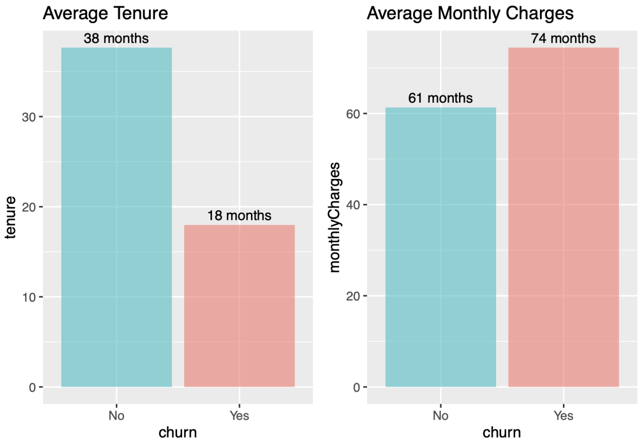


Exhibit 7

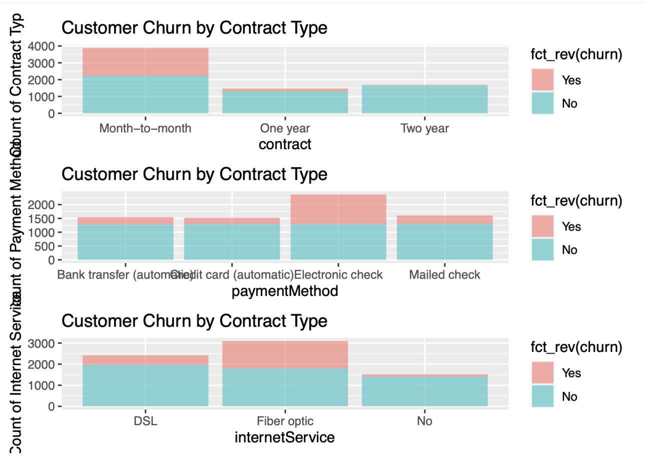


Exhibit 8

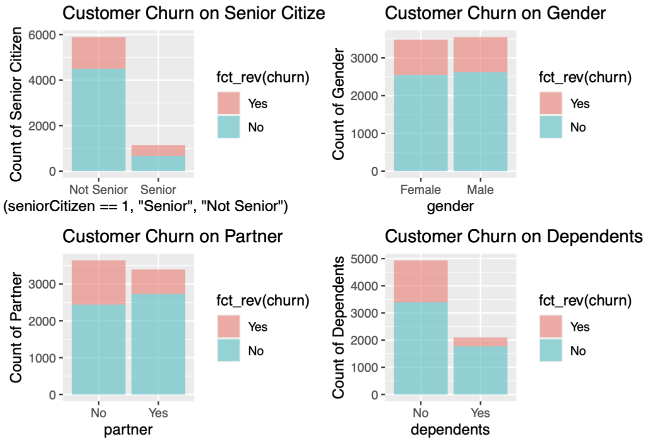


Exhibit 9

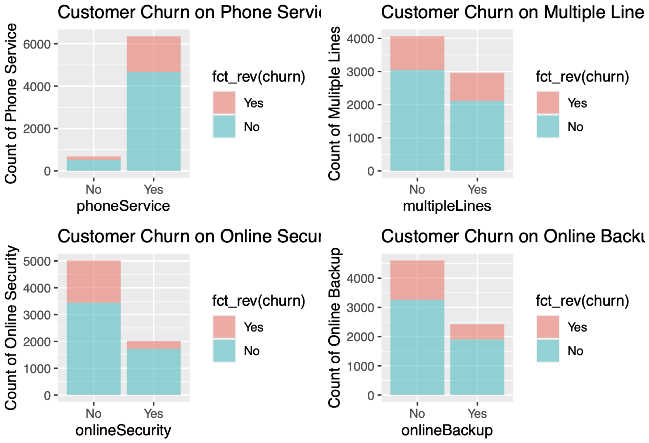


Exhibit 10

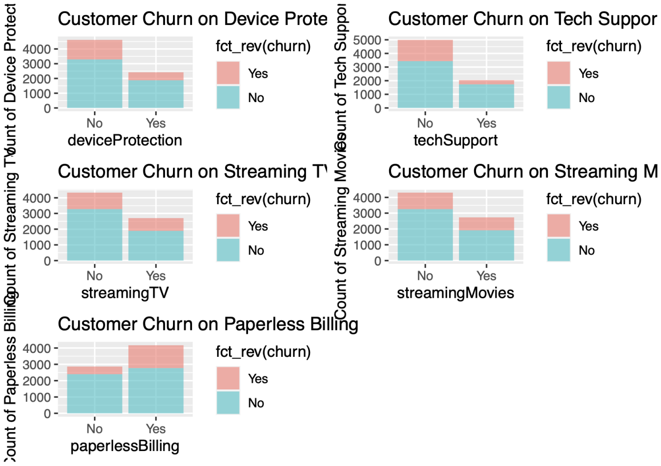


Exhibit 11

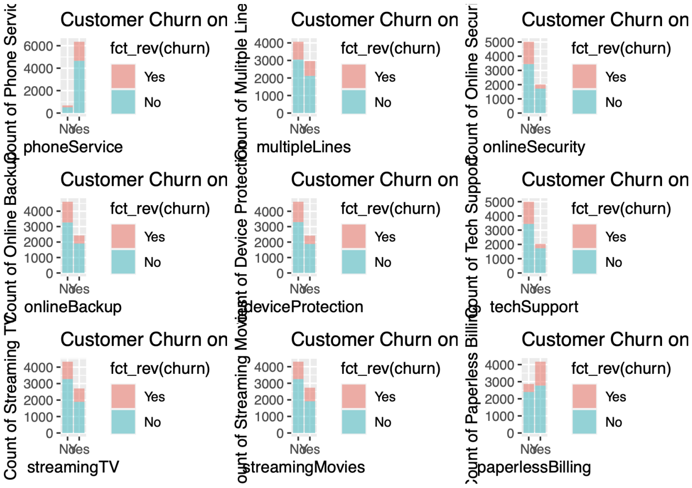


Exhibit 12

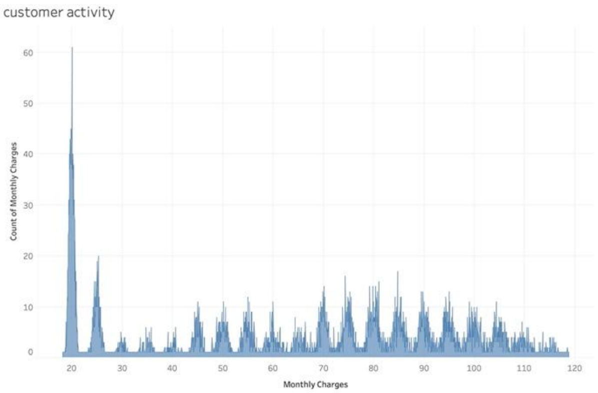


Exhibit 13

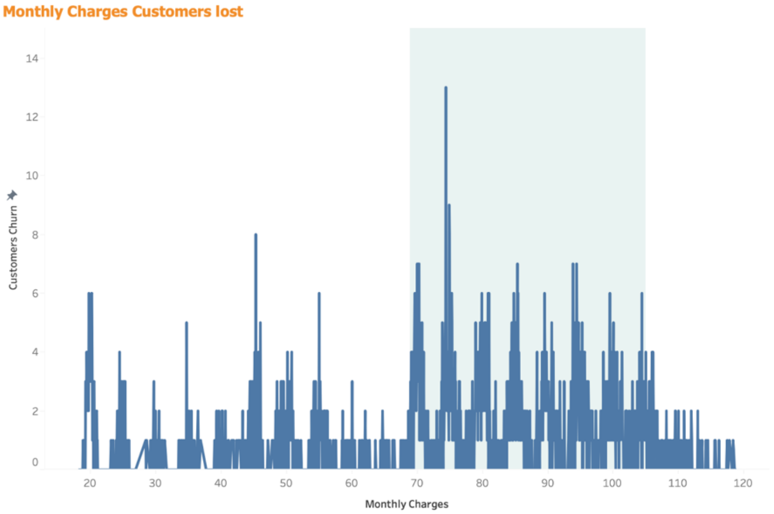


Exhibit 14

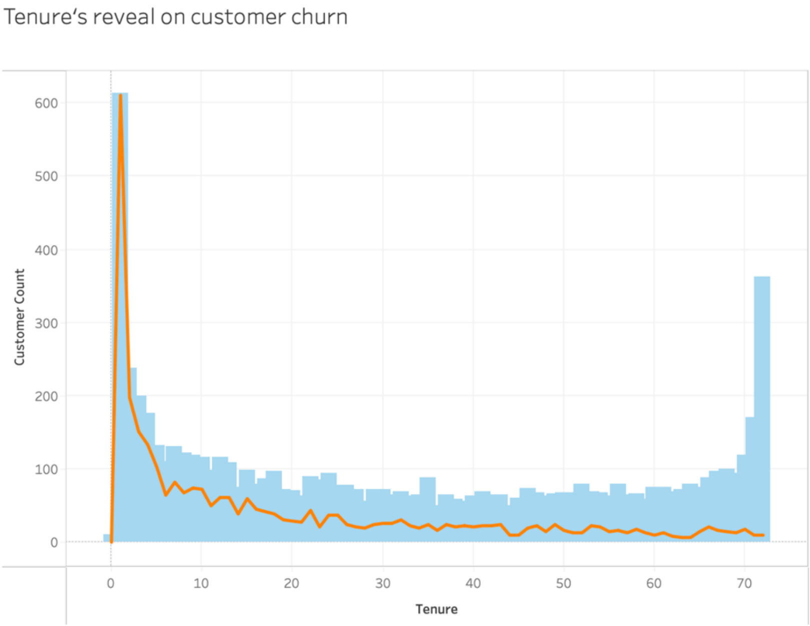


Exhibit 15

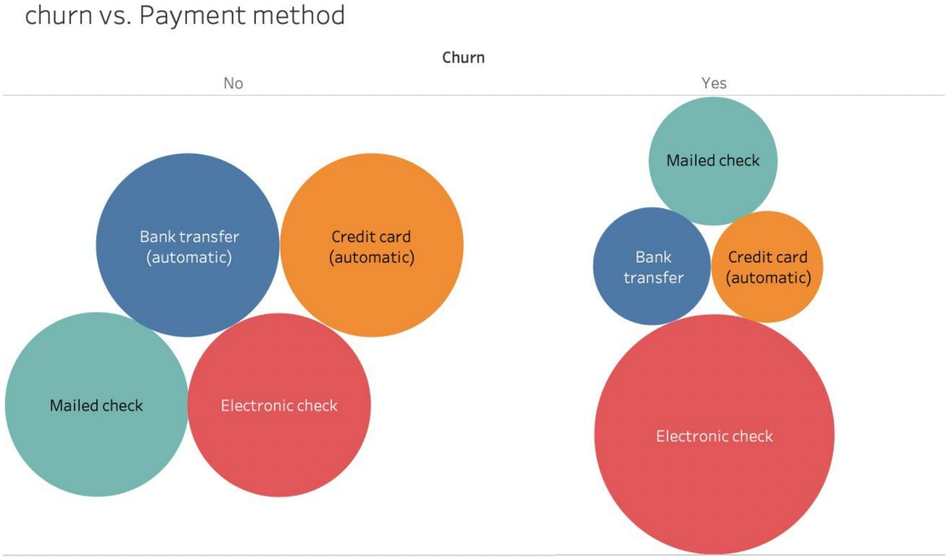


Exhibit 16

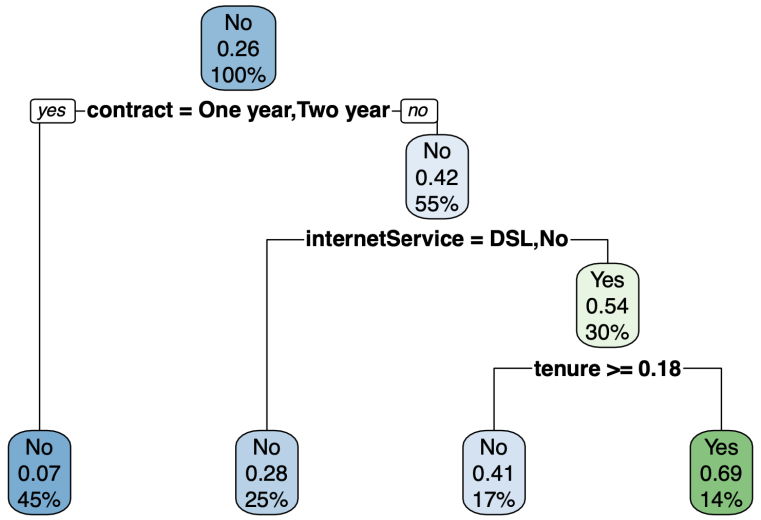
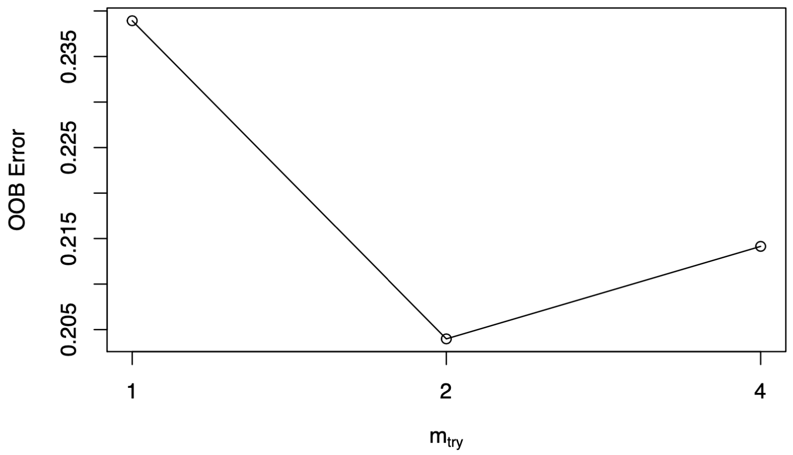
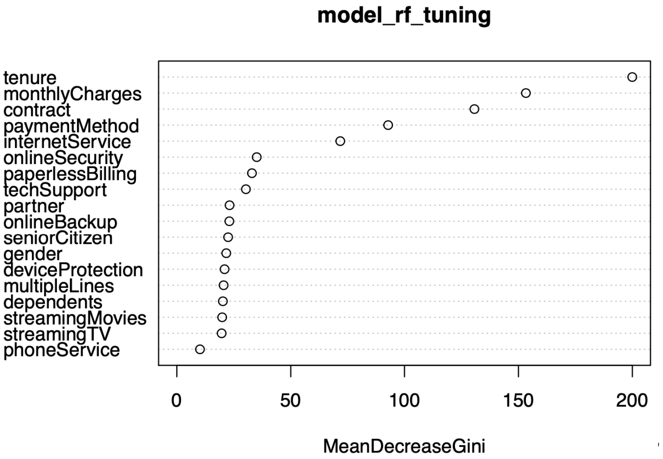


Exhibit 17



We

Exhibit 18


model_rf_tuning
tenure
monthlyCharges
contract
paymentMethod
internetService
onlineSecurity
paperlessBilling
techSupport
partner
onlineBackup
seniorCitizen
gender
deviceProtection
multipleLines
dependents
streamingMovies
streamingTV
phoneService
0
50
100
150
200
MeanDecreaseGini