Predicting Churn of Telecom Customers

The dataset

The dataset was obtained from Kaggle (https://www.kaggle.com/blastchar/telco-customer-churn). Each row contains information of a customer, including whether the customer left the program (churn) last month, which is the predicting target of the model of this paper. The dataset also contains 7043 rows and 20 features including gender, tenure, contract type, monthly charges, etc. A list of explanation of the features is in Appendix A.

Preprocessing

After using sapply() to see how many NAs in each column, we can see there are 11 rows where total charges are missing. These rows are omitted, considering 11 is a small number compared to 7043.

After observing the data, the following preprocessing steps are completed:

1. Change ‘SeniorCitizen’ column to factors from integers.
2. Remove the ‘cutomerID’ column, as it is not useful in our case.
3. Change ‘No phone service’ to ‘No’ for column ‘MultipleLines’
4. Change ‘No internet service’ to ‘No’ for six columns which have it.
5. Convert all columns with character variables to factors.

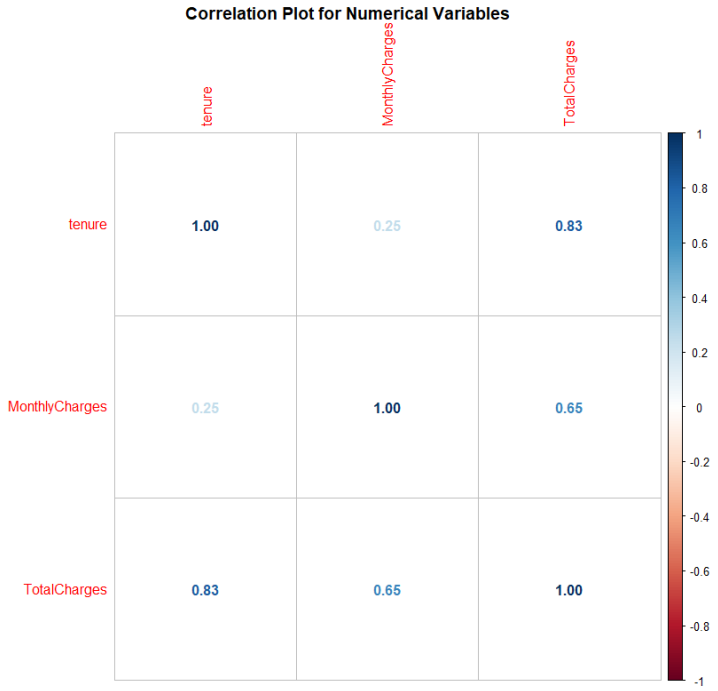
The structure of data frame is in as follows.

```
'data.frame':	7032 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

As it indicates the data set is clean and ready to be analyzed and modeled.

Exploratory data analysis

Correlation between the 3 numerical columns:



As the monthly charges and total charges are highly correlated, we delete the total charges column.

Random Forest

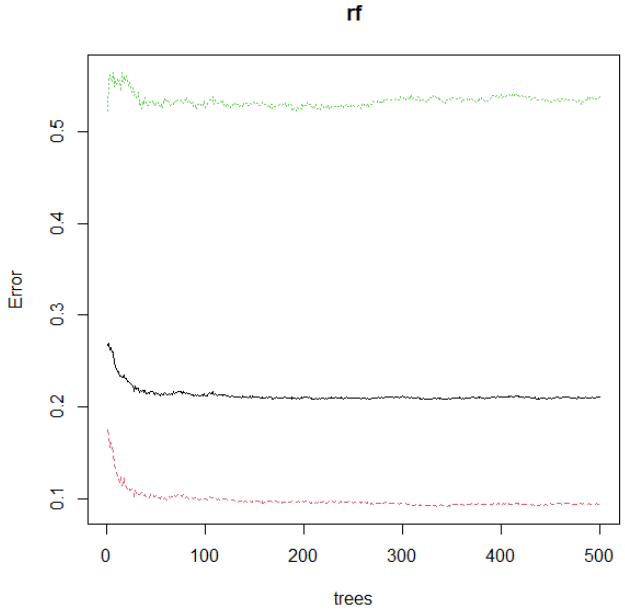
The data set is split into training and testing randomly by 7:3. A random forest classification model is fitted to the training set using default settings. We use random forest for this case, as the model can handle both categorical and numerical variables, and the dependent variable is a binary categorical variable. Random forest is based on bagging and ensemble learning. Compared to decision trees, it mitigates overfitting problems and reduces variance, and therefore performs better in terms of accuracy.

The initial random forest has an out-of-bag (OOB) error rate of 21.13%, and the following confusion matrix:

The initial random forest is used for the testing set and here is the result:

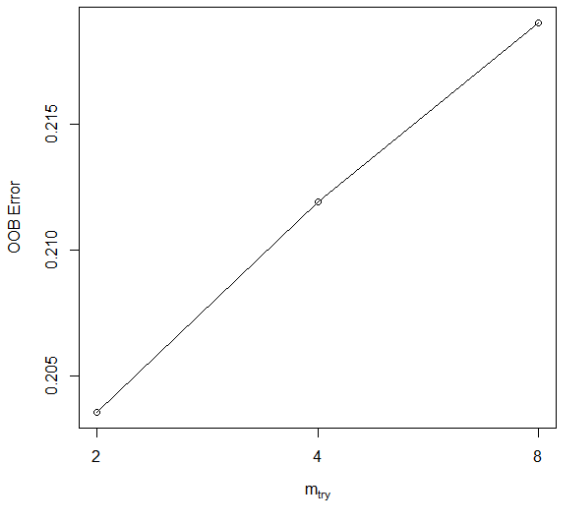
	No	Yes	Class. error
No	3284	345	0.0951
Yes	695	598	0.5375

Tuning



As the above figure indicates, the default setting of 500 trees is enough for the model, so 500 trees would be used in the tuning process.

Using the tuneRF() function, the following plot of OOB error against m_{try} is obtained:



When the number of variables tried at each split is set at 2, the OOB error is the smallest. Therefore, m_{try} is tuned to 2. A new random forest is then fitted for the same training set with the following result:

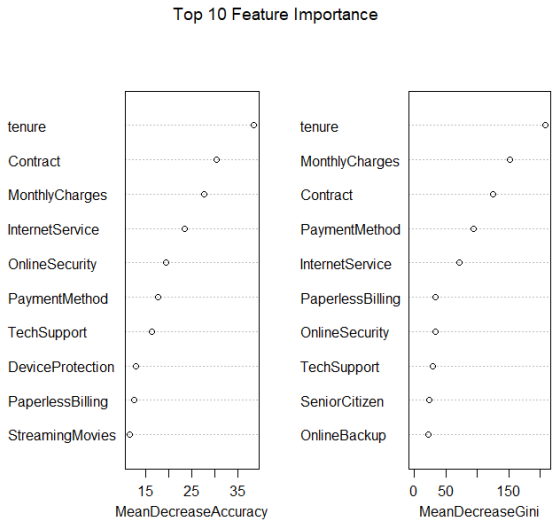
	No	Yes	Class. error
No	3361	268	0.0738
Yes	738	555	0.5708

The OOB error rate is reduced to 20.44% from 21.13%. And the model is tested using the testing set, with the result:

	No	Yes
No	1427	319
Yes	107	257

The prediction error rate is (107+319)/2110 = 20.19%

The feature importance plot is obtained. Tenure, contract type, monthly charges seem to be the most important features and gender does not seem to have relationship with churn. This can be used to help management decide about strategies to mitigate customer churns.



Notably, although the tuned model has a lower OOB error rate, it has a lower True Positive Rate (recall) than the original one. The telecommunication company might want to have a higher TPR with some compromise of overall accuracy, to better detect potential customers churning and response to keep the customers using discount or other strategies. If we use m_{try} = 8, the TPR will be even lower. Therefore, the actual tuning will depend on the company's actual requirements.

Reference

BlastChar. (2018, February 23). *Telco customer churn*. Kaggle. Retrieved March 13, 2022, from https://www.kaggle.com/blastchar/telco-customer-churn

Appendix A - Features Explanation

customerID
gender (female, male)
SeniorCitizen (Whether the customer is a senior citizen or not (1, 0))
Partner (Whether the customer has a partner or not (Yes, No))
Dependents (Whether the customer has dependents or not (Yes, No))
tenure (Number of months the customer has stayed with the company)
PhoneService (Whether the customer has a phone service or not (Yes, No))
MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service)
InternetService (Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup (Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport (Whether the customer has tech support or not (Yes, No, No internet service)
streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service)
streamingMovies (Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract (The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling (Whether the customer has paperless billing or not (Yes, No))
PaymentMethod (The customers payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic)))
MonthlyCharges (The amount charged to the customer monthly - numeric)
TotalCharges (The total amount charged to the customer - numeric)
Churn (Whether the customer churned or not (Yes or No))