Random Forest Project

Jacob Polomsky

June 30, 2024

A1. What are the major factors that affect Customer Tenure using Random Forest Regression?

A2. My goal for this analysis is to is to see factors that affect a customer Tenure without customer churn.

B1. A random forest is an ensemble of decision trees. Each tree is made a little different from the last, like some data points might be resampled, so the model, then after all the trees are made, the model aggregates all of the predictions for all the trees. Then, the model returns the average of all of the predicted values. Random forest regression can help predict complex relationships like tenure, and aggregating the result of the trees reduces the chance of overfitting.

B2. One assumption with Random Forest is that there are no formal distributions because it is a non-parametric model. This allows it to handle complex relationships because it doesn't adhere to formal distributions. The model has difficulties in following a formal distribution. It makes the relationship more complex and harder to interpret and can cause possible overfitting than with another model.

B3. The packages I used are Pandas to store and manipulate the data and NumPy to do math functions like square root. Scikit-Learn will run my regression model and check its accuracy.

C1. My goal is to manipulate the data to be usable for running the random forest model, first by checking for duplicates, null values, and outliers and dropping those values if they exist. Next is to convert categorical variables to numerical variables by converting them into dummy variables. Then, I split my data set into X and Y.

C2. The Variable I used two types of variables: categorical and numerical variables. The categorical variables are State, Area, Martial, Gender, Internet Service, and Contract. The numerical variables are Age, income, children, outages per week, email, contracts, monthly charge, bandwidth, and tenure.

C3. The steps I used for this analysis is to first check for duplicates which there were none.

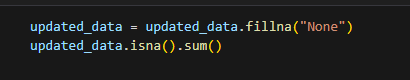
A screen shot of a computer

Description automatically generated

Second was to check for null values which there were 2000 in Internet Service. I noticed after looking at the csv that the values were not NA, but should have been None. So, I replaced Na values with None.

A screenshot of a computer

Description automatically generated



Then I checked for outliers in the data and found none

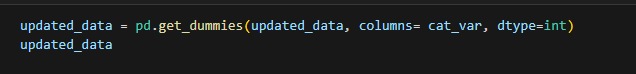
A screenshot of a computer

Description automatically generated

A screenshot of a computer

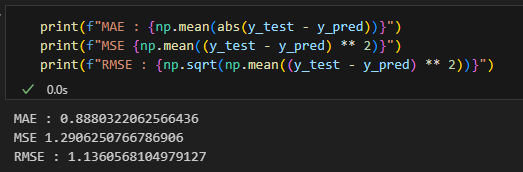
Description automatically generated

I then converted Categorical variables into dummy variables, since the model is nonlinear I did not need to drop the first.



C4. Updated\_data.csv

D1. The four csv are attached.

D2. I split my X and Y variables into train and split sets. I then ran a default random forest regression model with no hyperparameter tuning. I then fit the Train set to the random forest model. I then predicted the values for the Train and test set, the train set was unnecessary want to check for overfitting. I did not run the intermediate calculations in my code; I ran the functions, but here is a screenshot of how you would calculate the accuracy scores.

D3. RF\_model.ipbny

E1. The way I am going to check for the model accuracy is by using mean absolute error or MAE, which subtracts the actual value from the predicted value; this is called residual, then takes the absolute value of all residuals, then takes the mean of all those absolute residuals. Mean squared error or MSE takes the residual, then squares the values, and then takes the mean of all the residual squared. Root means squared error, or it takes the square root of MSE. R-squared measures how well the model fits the data.

E2. The accuracy for our test set model is an MAE of 0.87, an MSE of 1.29, and an RMSE of 1.13. These scores show that the model is off by a month on average for every prediction. When you look at the standard deviation, which is 26 months, it shows the model has a good grasp on how long a customer stays. The R-squared score is 0.998. This means the model can explain 99.8% of the variance in the data, implying that the model can accurately predict why a value went higher or lower. This model does an excellent job of accurately predicting customer tenure with the company.

A screenshot of a computer

Description automatically generated

E3. A limitation in this analysis is that the Train set performs better, insinuating that there might be some overfitting, but even the test set performs well, so I do not think that is a major concern. I ran the data a couple of times, and the results were similar each time with similar accuracy scores, but this model does not have multiple cross-validations. This model might perform well only on this test set and not on others. So, introducing more data in multiple cross-validation always helps to see how accurate your model is.

E4. This model should always be tested and updated with new data. With how well this model performed on this Train and test set, this model is very accurate in predicting customer tenure. I would recommend its use to see how long you predict a customer will stay with the company. Possible in gauging how long a new customer will stay.