Part 1: Research Question

1. Can a customer's monthly charges be predicted using a random forest regression?

2. This analysis aims to predict a customer's monthly charges based on demographics and service information. Ensuring that the company offers rates that are reasonable and competitive with other providers can help with retention.

Part II: Method Justification

1. Random Forest regression is an ensemble method that uses multiple decision trees to make predictions. Each decision tree uses a data sample for the original set without replacement. This technique is known as bootstrapping. The predictions of the decision trees are averaged into a final prediction. (IBM Cloud Education, 2020) The expected outcomes are an accuracy score, in this case, an R-squared score and a mean squared error score.

2. Random Forest regression is a non-parametric model. It does not assume anything about the distribution of the data. It can process skewed and multi-modal data. (Mendekar, 2021)

3. Six packages and libraries are used for this analysis. Pandas and NumPy are used to import the data set and for data preparation. Train_test_split, RandomForestRegressor, mean_squared_error, and r2_score were imported from sklearn. Train_test_split created the training and testing data sets. RandomForestRegressor built the model. Mean_squared_error and r2_score scored the accuracy of the model.

Part IV: Data Preparation

1. Sklearn requires all categorical variables to be converted into numeric values. Dummy variables are created for this purpose using get_dummies from the pandas library. All values become one if the record falls into that category. For example, if the customer has paperless billing, a one will be recorded in the PaperlessBilling_Yes column.

2. The target variable in this analysis is MonthlyCharge which is a continuous variable. All other variables are the predictor variables. They are:

Continuous: Population, Children, Age, Income, Outage_sec_perweek, Email, Contracts, Yearly_equip_failure, Tenure, Bandwidth_GB_Year

Categorical: Churn, Area, Marital, Gender, Techie, Contract, Port_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, Streaming_TV, StreamingMovies, PaperlessBilling, PaymentMethod

3. The data was checked for any null values with df.isnull().values.any(). The outcome was False, so there are no null values. Next, df.duplicated().values.any() checked for duplicate rows. The

outcome was False again. Columns that are not being used for this analysis are dropped. Finally, dummy variables are created for the categorical variables using pd.get_dummies. The original column was dropped, and the dummy variables were added using df.drop() and df.merge().

4. See attached file: predictive_prepared_churn.xlsx

Part V: Analysis

1. See attached files: X_train.xlsx, X_test.xlsx, y_train.xlsx, y_test.xlsx

2. After splitting the data set into training and testing, the RandomForestRegressor was fitted to the training data. The X_test data set is used to make predictions. The accuracy was calculated with an R-squared score. The mean squared error and the root mean error are also calculated.

```
Test set R squared score: 0.999990
Test set Mean Squared Error: 0.018317
Test set Root Mean Squared Error: 0.135339
```

3. See attached file: D209 Task 2 Predictive Analysis.pdf

Part V: Data Summary and Implications

1. The performance of the model is measured using R-squared and the Mean Squared Error. The R-squared is 0.99, which is close to perfect. The Mean Squared Error is tiny at 0.018. The model's predictions come very close to the actual value.

2. The R-squared is very high. A score of one is perfect, and the model has a score of 0.99. The model does a near-perfect job of predicting the monthly charge. The Mean Squared Error is minimal at 0.018. There is very little difference between a predictive point and an actual point.

3. One limitation of using random forest regression to predict is that the model does not generalize well. It can only make predictions within the range of the training data. (Thompson, 2019)

4. The company could give current, and potential customers access to the model for predicting their monthly bills. The customers could use it to see which features fit within their budget. Potential customers could use it to compare rates with other companies.

Works Cited

IBM Cloud Education. (2020, December 7). *Random Forest*. Retrieved Aug 2021, from IBM Cloud Learn Hub: https://www.ibm.com/cloud/learn/random-forest

Mendekar, V. (2021, Feb). *Machie Learning - it's all about assumptions*. Retrieved August 2021, from KDNuggets: https://www.kdnuggets.com/2021/02/machine-learning-assumptions.html

Thompson, B. (2019, December 17). *A limitation of Random Forest Regression*. Retrieved Aug 2021, from towards data science: https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f