Part I: Research Question

A1. Proposal of Question

Using the K-nearest neighbor model, can we accurately predict customer churn based on a set of predictor variables within the given dataset to understand each of those variables' effect on customer churn and to reduce customer churn?

A2. Defined Goal

The goal and objective of this analysis are to use K-Nearest Neighbor (KNN) to help us further understand why customers churn and what variables affect their decision to churn. By understanding these variables, we hope to predict whether a customer will churn given a certain threshold of K-Value. For our analysis, we will be using the predictor variables “Sum_services”, “Area”, “Employment”, etc against our target variable “Churn”.

Part II: Method Justification

B1. Explanation of Classification Method

The use of KNN for our purpose will provide us with a good prediction of whether a customer will churn based on its “feature similarity”. That is to say, given a new data point, how well can the model predict whether a customer will churn based on how close its “feature” matches to the points around it in the training data set (*KNN Algorithm…*, 2022). The expected outcome is that if its “feature similarity” is closest to churn, it will classify the data point correctly.

B2. Summary of Method Assumption

One of the assumptions of a KNN model is that data points that exist close to one another are highly similar, while if the data points are far from another group, it is dissimilar to those data points (Nelson, 2020). This assumption summarizes that given its "feature similarity", data points that are close to each other in the model should be consider similar while data points that are farther away from one another, should be considered different.

B3. Packages or Libraries List

For this analysis, we will be using Python as it has a wide array of libraries and packages that is suited for our purpose. Some of the packages and libraries necessary are numpy, pandas, scikit-learn, and seaborne among other libraries. The versatility of Python and its relationship to statistics and machine learning also make it a perfect fit for our use in this analysis (Talab, Z., 2021). Python will allow us to prepare our data during the data cleaning process using numpy and pandas. We can then use it to further plot our data during the exploratory analysis using libraries like seaborne to visualize our predictor and response variables and to detect outliers. Finally, Python with its diverse libraries will allow us to perform our necessary predictive analysis during our data modeling phase using scikit-learn. A list of packages or libraries used for this analysis was also provided below:

- Pandas
- Numpy
- Scipy
- Matplotlib
- Seaborn
- Sklearn

Part III: Data Preparation

C1. Data Preprocessing

One of the data preprocessing goals for this classification method is to make sure that we scale our continuous variables to provide an accurate measure for our KNN model (Roy, 2021). This step is important as it allows the classification analysis to predict variables more correctly with huge differences in numbers such as “Income” and “Reliability”. Performing this task will allow us to predict customer churn more accurately.

C2. Data Set Variables

In our raw data, we have a total of 10,000 observation rows with 52 variable columns. After the data cleaning process, our dataset was reduced to 9978 observations with 1 response variable, 12 predictor variables of which were later expanded to 21 after creating dummy variables. Since our research question is to understand and predict churn, our target variable is “Churn”. Our predictor variables are “Area”, “Employment”, “Income”, “Contract”, “Sum_services”, “Timely_Fixes”, “Timely_Replacement”, “Reliability”, “Options”, “Respectable_Response”, “Courteous_Exchange”, and “Evidence_of_active_listening”. Before we do our initial analysis, we will create dummy variables for our categorical variables “Area”, “Employment, “Contract”. Since K-Nearest Neighbor can handle raw dummy variables, we will not be removing one level (*Categorical Predictors: How Many Dummies to Use in Regression vs. k-Nearest Neighbors*, 2015). These new dummy variables will be “Dummy_Rural”, “Dummy_Urban”, “Dummy_Suburban”, “Dummy_Full Time”, “Dummy_Part Time”, “Dummy_Retired”, “Dummy_Student”, “Dummy_Unemployed”, “Dummy_Month-to-month”, “Dummy_One year”, “Dummy_Two Year”. Below is the list of which variables are continuous and categorical:

Categorical Variable:

- Area, Employment, Contract, Churn, Dummy_Urban, Dummy_Suburban, Dummy_Part Time, Dummy_Retired, Dummy_Student, Dummy_Unemployed, Dummy_One year, Dummy_Two Year, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies.

Continuous Variable:

- Income, Timely_Fixes, Timely_Response, Timely_Replacement, Reliability, Options, Respectable_Response, Courteous_Exchange, and Evidence_of_active_listening.

C3. Steps for Analysis

To prepare the data for analysis, we will:

1. Import all necessary libraries and datasets into python.
2. Identify the unique entries for each of the variables that we are working with.
3. Drop all unused/duplicate columns, impute missing data/remove them, and remove duplicate rows.
4. Recode all binary categorical columns to 1 and 0.
5. Recode categorical columns with more than two levels to an ordinal variable 0, 1, 2, for use on later analysis. These will be later recoded to dummy variables.
6. Update the name of all the Customer Survey columns using the key names provided by the course document (1=Timely_Response, 2=Timely_Fixes, 3=Timely_Replacement, etc)
7. Find outliers by creating a Z-Score or identifying them through univariate/bivariate analysis.

8. Create dummy variables for categorical data that have more than two levels for our KNN classification model.

Part IV: Data Summary and Implications

E1. Accuracy and AUC

Using KNN, we were able to identify that the model's accuracy was initially 67% accurate at predicting the correct result. Predictor variables were then scaled and improved the model to 70% accuracy. Hyper-tuning was then implemented which increased the model accuracy to 73%. Using this final model, we were able to obtain a 74% precision score (when we predict yes, and the result was yes), 97% specificity score (when we predict no, and the result was no), and an AUC score of 71% (explained variance within the dataset) (Sharma, 2021). This indicates that our model was acceptable, at 70% accuracy, in predicting customer churn.

The AUC score of the model was 71%, indicating that it is an acceptable discrimination at correctly classifying a customer (Draelos, 2020). That is, it is 71% accurate at correctly classifying when a customer has or has not churned and was able to explain 71% of the variance within the dataset.

E2. Result and Implications

Given an acceptable accuracy prediction score of 73% and an AUC score of 71%, the implications of this are we can conclude that the model is acceptable at accurately predicting or classifying customer churn. There were some other notable results such as the model indicating an average precision score of 74%. Another value that stood out was our specificity score at 97%, which may indicate that this model can be used to identify which variables retain customers and slow customer churn. We did perform various tasks to improve the model such as hyper-tuning with the best parameter but it was only

marginally effective at increasing our accuracy score. With these values, it may provide us some insight into what affects customer churn.

E3. Limitations

One of the limitations of the data analysis may have had to do with the number of predictor variables. We have 21 predictor variables and through our work on a different classification model with this same dataset, we did identify some predictor variables to not be statistically significant and since we did not remove those variables for this model, it may have affected our calculated accuracy and AUC score. In our next model, we may need to reiterate with either more variables, reduce variables, or implement a feature selection such as Step-wise Forward or Backward Selection. This can potentially increase the accuracy of the model and provide a clearer picture of how effective the model is.

E4. Course of Action

Based on our results (accuracy score = 73%, AUC score = 71%), we understood that the model was acceptable at correctly predicting and classifying customer churn. Therefore per our research question, we could acceptably predict customer churn using K-Nearest neighbor and achieve some understanding of variables that affects customer churn. The next course of action will be to apply the model to new data and see if we can predict customer churn to reasonably understand how effective our model is and how we can improve upon our understanding of what makes a customer churn. It is our eventual goal that we can reduce customer churn rate by predicting which variables ultimately affect it and improve our overall business application.

Part V. Demonstration

F. Panopto Recording

G. Sources for Third-Party Code

Draelos, V. A. P. B. R. (2020, February 2). *Measuring Performance: AUC (AUROC)*. Glass Box. https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/#:%7E:text=1%20An%20AUROC%20of%200.5%20%28area%20under%20the,the%20figure%20above%29%20corresponds%20to%20a%20perfect%20classifier

F. (2020, August 15). *Scikit-Learn Pipeline Examples Last updated: 15 Aug 2020*. Queirozf.Com. https://queirozf.com/entries/scikit-learn-pipeline-examples

Fine-tuning your model - Hyperparameter tuning. (2022). DataCamp. https://campus.datacamp.com/courses/supervised-learning-with-scikit-learn/fine-tuning-your-model?ex=9

Kumar, V. (2021, September 4). *KNN Classifier in Sklearn using GridSearchCV with Example*. MLK - Machine Learning Knowledge. https://machinelearningknowledge.ai/knn-classifier-in-sklearn-using-gridsearchcv-with-example/

Radečić, D. (2021, December 23). *ROC and AUC — How to Evaluate Machine Learning Models in No Time*. Medium. https://towardsdatascience.com/roc-and-auc-how-to-evaluate-machine-learning-models-in-no-time-fb2304c83a7f

Sharma, P. (2021, December 11). *Decoding the Confusion Matrix - Towards Data Science*. Medium. https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

Split data into testing and training and convert to csv or excel files. (2020, July 23). Stack Overflow. https://stackoverflow.com/questions/63054593/split-data-into-testing-and-training-and-convert-to-csv-or-excel-files

Sun, Q. (2018, May 19). *How to deal with Cross-Validation based on KNN algorithm, Compute AUC based on Naive Bayes algorithm*. Medium. https://medium.com/@svanillasun/how-to-deal-with-cross-validation-based-on-knn-algorithm-compute-auc-based-on-naive-bayes-ff4b8284cff4

H. Sources

Categorical predictors: how many dummies to use in regression vs. k-nearest neighbors. (2015). BzST. http://www.bzst.com/2015/08/categorical-predictors-how-many-dummies.html

KNN Algorithm In Machine Learning | KNN Algorithm Using Python | K Nearest Neighbor | Simplilearn. (2018, June 6). [Video]. YouTube. https://www.youtube.com/watch?v=4HKqjENq9OU

KNN Algorithm - Finding Nearest Neighbors. (2022). Www.Tutorialspoint.Com. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

Kohli, S. (2021, December 12). *Understanding a Classification Report For Your Machine Learning Model*. Medium. https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Nelson, D. (2020, August 24). *What is a KNN (K-Nearest Neighbors)?* Unite.AI. https://www.unite.ai/what-is-k-nearest-neighbors/

Roy, B. (2021, December 14). *All about Feature Scaling - Towards Data Science*. Medium. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Talab, Z. (2021, November 12). *Benefits of Python for AI*. Developer.Com. https://www.developer.com/languages/benefits-of-python-for-ai/#:%7E:text=In%20addition%20to%20its%20versatility%2C%20Python%20is%20a,pl

atforms%3B%20Windows%2C%20MacOS%2C%20Linux%2C%20Unix%2C%20and%20many%20more.

Z. (2021, September 9). *What is Considered a Good AUC Score?* Statology. https://www.statology.org/what-is-a-good-auc-score/