Part I: Research Question

A. Describe the purpose of this data mining report by doing the following:

1. **Propose one question relevant to a real-world organizational situation that you will answer using one of the following classification methods:**
   - ***k*-nearest neighbor (KNN)**
   - **Naive Bayes**

   Which customers churned given other variables in our data set? I will use k-nearest neighbor classification to answer this question.

2. **Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.**

   My goal is to build a KNN model that predicts if a customer churned or not based on other variables in an observation.

Part II: Method Justification

B. Explain the reasons for your chosen classification method from part A1 by doing the following:

1. **Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.**

   The KNN algorithm measures the distance between the target and features of an observation. It saves these distances in an array. When you pick a number, k, of nearest neighbors and run the classification, it uses the mode of the k labels to predict an observation's label. These labels, along with scores from a classification report, are my expected outcomes. (Harrison, 2018)

2. **Summarize one assumption of the chosen classification method.**

   One assumption of the KNN method is that datapoints that are close in proximity will be alike in nature. If this assumption were untrue, then the KNN algorithm would not be useful. (Harrison, 2018)

3. **List the packages or libraries you have chosen for Python or R and justify how *each* item on the list supports the analysis.**

   I've chosen Python as my language. Libraries and packages include:
   - Pandas to import/export and clean the data
   - Numpy to create arrays
   - Matplotlib.pyplot to create visualizations
   - sklearn.model_selection: train_test_split to split the data into training and testing data
   - sklearn.preprocessing: StandardScaler to normalize the feature data
   - sklearn.neighbors: KNeighborsClassifier to create the regression model
   - sklearn.metrics: classification_report, confusion_matrix, roc_auc_score to evaluate the model

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:

1. **Describe one data preprocessing goal relevant to the classification method from part A1.**
   Change all object type data to numeric data so that it can be passed into the KNN classification model.

2. **Identify the initial data set variables that you will use to perform the analysis for the classification question from part A1 and classify *each* variable as continuous or categorical.**
   Variables include:
   - CaseOrder: continuous
   - State: categorical
   - Zip: continuous
   - Lat: continuous
   - Lng: continuous
   - Population: continuous
   - Area: categorical
   - TimeZone: categorical
   - Children: continuous
   - Age: continuous
   - Income: continuous
   - Marital: categorical
   - Gender: categorical
   - Outage_sec_perweek: continuous
   - Email: continuous
   - Contacts: continuous
   - Yearly_equip_failure: continuous
   - Contract: categorical
   - InternetService: categorical
   - PaymentMethod: categorical
   - Tenure: continuous
   - MonthlyCharge: continuous
   - Bandwidth_GB_Year: continuous
   - Item1-Item8: categorical
   - Churn_Yes: categorical
   - Techie_yes: categorical
   - Port_modem_Yes: categorical
   - Tablet_Yes: categorical
   - Phone_Yes: categorical
   - Multiple_Yes: categorical
   - OnlineSecurity_Yes:categorical

- OnlineBackup_Yes: categorical
- DeviceProtection_Yes: categorical
- TechSupport_Yes: categorical
- StreamingTV_Yes: categorical
- StreamingMovies_Yes: categorical
- PaperlessBilling_Yes: categorical

3. **Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.**
   I will first import the original dataset and all necessary libraries. I will drop the variables Customer_id, Interaction, UID, and Job because they are not necessary from a business perspective. I will drop City and County because they are redundant with Zip. Then I will change object type categorical data to category codes, then create dummy variables for object type categorical data that have only "Yes" and "No" values. Then I will drop all the "No" dummy variables to eliminate redundancy. See the Keim D209 Task One Code document for the code segments.

4. **Provide a copy of the cleaned data set.**
   See Keim D209 Task One Clean Data document for the cleaned data set.

Part IV: Analysis

D. Perform the data analysis and report on the results by doing the following:

1. **Split the data into training and test data sets and provide the file(s).**
   See Keim D209 Task One Training Data and Keim D209 Task One Testing Data documents for the training and testing data sets.

2. **Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.**
   After splitting the data into 80% training and 20% testing data, I created feature and target arrays for both sets. Then I scaled all of the features. Using a for loop, I plotted the accuracy scores for the testing and training data to determine the best value of K for my model, which was 17. I then called the classifier with a K of 17, fir the model to the training set, and predicted labels with the testing set. Then I evaluated the model using a confusion matrix, classification report and AUC score. All calculations and code can be found in the Keim D209 Task One Code document.

3. **Provide the code used to perform the classification analysis from part D2.**
   See Keim D209 Task One Code document.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:

1. **Explain the accuracy and the area under the curve (AUC) of your classification model.**

   The accuracy of my model is 0.74, meaning it correctly predicts that a customer churned 74% of the time. The AUC is 0.59, meaning that it correctly predicts the label, either 'churned' or 'did not churn' 59% of the time.

2. **Discuss the results and implications of your classification analysis.**

   'Did not churn' has a precision score of 0.78, and a recall score of 0.92. This means that the model has an accuracy of 78% when predicting that customers will stay with the company and correctly identified 92% of all customers who stayed with the company.

   'Did churn' has a precision score of 0.52 and a recall score of 0.26. This means that the model has an accuracy of 52% when predicting that customers will leave the company and correctly identified 26% of all customers who left the company. (Kohli, 2019)

3. **Discuss one limitation of your data analysis.**

   While the analysis can help us predict what types of customers will stay with the company, it cannot do this with 100% accuracy, and it also cannot tell us why customers are choosing to stay or go.

4. **Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2**

   The scores discussed in part E2 indicate that the model does a better job predicting who will stay with the company than predicting who will leave the company. I recommend using the model to examine the features of customers predicted to stay to maximize customer retention.

Bibliography

Harrison, O. (2018, 09 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm*. Retrieved from towards data science: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Kohli, S. (2019, 11 17). *Understanding a Classification Report For Your Machine Learning Model*. Retrieved from Medium: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

No third party code was used.