Part I: Research Question

1. Which customers are more likely to terminate services (Churn) with the telecommunication company?

2. Using K-nearest neighbor, this analysis aims to predict which customers will terminate services, allowing the company to take targeted preventive measures.

Part II: Method Justification

1. K-nearest neighbor predicts the label of a data point by looking at the 'k' closest points where 'k' is a predetermined number. The data point is then labeled according to which classification has the greatest number of points closest to the point of inquiry. (Bowne-Anderson, n.d.) The analysis outcome is a score that measures the model's accuracy and a ROC curve that shows how well the model distinguishes between churn and not churn. (Narkhede, 2018)

2. A K-nearest neighbor model assumes that similar data points are close to each other. The knowledge gained from one point can be used to draw conclusions about other points. (Grant, 2019)

3. Eleven packages and libraries were used for this analysis. Pandas and NumPy were used to import the dataset and for data preparation. Matplotlib was used to create the roc graph. Preprocessing, StandardScaler, KNeighborsClassifier, train_test_split, metrics, roc_curve, make_pipeline, and confusion_matrix were imported from sklearn. These packages scaled the data, split the data, created the model, and measured how well the model worked.

Part III: Data Preparation

1. Sklearn requires that all the predictive variables be continuous. Dummy variables are created to fulfill this requirement. Dummy variables take categorical data and use a numeric value in its place. For this analysis, 1 and 0 are used. One indicates a positive for that category.

2. The following variables were used to build this model:

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

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

3. The data was checked for null values using df.isnull().values.any(). Df.duplicated().values.any() checked for any duplicate rows. To drop the unnecessary columns, df.drop() was used. Dummy variables were created using pd.get_dummies(). The extra columns

were dropped with df2.drop(). Finally, a copy of the prepared data set was saved using df2.to_csv().

4. Attached as classification_prepared_churn.csv

Part IV: Analysis

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

2. The data set was standardized using a pipeline. The values between the different variables can vary greatly. For example, Children has a range from 0 to 10. Income has a range from 348 to 258900. Income would have a more significant influence over the outcome if the data set were used as-is. Next, the pipeline was fitted to the training data, and the X_test set was used to make predictions. A confusion matrix was created to see the distribution of predicted outcomes. A classification report was created to look at the precision, recall, and accuracy scores. The precision was 86%, and the recall was 91%.

	precision	recall	f1-score	support
0	0.86	0.91	0.89	2210
1	0.71	0.59	0.65	790
accuracy			0.83	3000
macro avg	0.79	0.75	0.77	3000
weighted avg	0.82	0.83	0.82	3000

3. Code attached with file D209 Classification Analysis Code.pdf

Part V: Data Summary and Implications

1. The accuracy of the model is 83%. The model correctly predicts that a person will churn 83% of the time. The AUC or area under the curve is 75%. The model can correctly classify true Churn_yes 75% of the time.

2. The accuracy and AUC scores are both high. This model does a good job of predicting which customers will terminate their services with the company. These results might be improved with hyperparameter tuning or feature selection.

3. One limitation of the analysis is the size of the data set. The data set represents a tiny portion of the total population. A quick Google search showed that three of the largest telecommunication companies have 20 million customers each. The data set with 10,000 records is not very large in comparison.

4. It is recommended that the company offer incentives to the customers predicted to turn. Offering things like free upgrades or lower locked-in rates to customers predicted to terminate could retain them and forego the costs of acquiring new customers.

Works Cited

Bowne-Anderson, H. (n.d.). *The classification challenge*. Retrieved July 2021, from Datacamp: https://campus.datacamp.com/courses/supervised-learning-with-scikit-learn/classification?ex=6

Grant, P. (2019, July 21). *Introducing k-Nearest Neighbors*. Retrieved July 2021, from towards data science: https://towardsdatascience.com/introducing-k-nearest-neighbors-7bcd10f938c5

Narkhede, S. (2018, June 26). *Understanding AUC-ROC Curve*. Retrieved Aug 2021, from towards data science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5