K-Nearest Neighbor Project

Jacob Polomsky

June 28, 2024

A1. What is the probability of Customer Churn being predicted using a K-nearest Neighbor Classification method?

A2. My goal in this analysis is to gain a better understanding of what variables correlate to Churn.

B1. How K-Nearest Neighbor, also called KNN, predicts outcome is by looking at the 'k' closest data points to the value you want to predict. It is then assigned to the Classification group with the most common or the majority party. KNN uses Euclidean distance for measuring distance the example below has the value you want to predict in black. It looks at the five closest data points, three blue and two red since there are more blue dots than red dots. This predicted data point would be blue.A diagram of a diagram

Description automatically generated

B2. One assumption with KNN is that features can be measured by distance. Since the model measures distance from one point to another, it is essential that the distances can be measured; otherwise, you can't even run the model, and if you do, the model will perform very poorly.

B3. The packages that I am using are Pandas and NumPy to store and manipulate the dataset. Scikit-Learn to run the machine learning algorithm, scale the data, and test the model's accuracy. Matplotlib will plot the confusion matrix and roc curve.

C1. My goal for data processing is to make my data usable and in the ideal format for the KNN model. First, by scaling my numeric variables, since KNN looks at a distance, but if some variables have a bigger variance between max value and min values, the distance will look bigger for the KNN method, making it run less effectively compared to scaled data. I checked for duplicates, null values, and outliers as they may throw off the model predictions. I replaced categorical variables with 1 on 0 for yes-no questions and dummy variables for variables that had more than one group.

C2 The variable I used both numerical and categorical variables in my dataset. The numeric variables are Age, Income, Monthly Charge, Tenure, Outages per week, Bandwidth per year, and Yearly equipment failure. The categorical variables are Gender, Contract, Area, Martial, Internet Service, Paperless Billing, and Churn.

C3. I scaled the data using the code below

A screen shot of a computer program

Description automatically generated

I check for duplicates with Code below and found 0 duplicates

A screen shot of a computer

Description automatically generated

I checked for null values with code below and found 2000 missing data points in Internet Service, but looking at the data set the NA values are supposed to be None. I used fillna to replace null values with None.

A screenshot of a computer

Description automatically generated

A black screen with white text

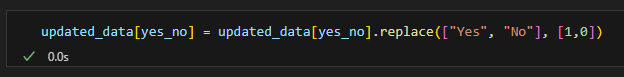
Description automatically generated

Used the Describe Function to look for Outliers and did not find any

A screenshot of a computer

Description automatically generated

I converted Yes/ NO questions into 1 or 0. With code below



Converted the rest of the categorical variable into dummy variables

A screen shot of a computer

Description automatically generated

C4. See updated\_data.csv

D1 see csv attached

D2. I split my X and Y variables into train and test splits. I then fit my Train and split set into the KNN model. 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, so I wanted 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.A screenshot of a computer program

Description automatically generated

D3. See KNN\_model.ipynb

E1. Accuracy measures how many you got right, precision measures how many you predicted were right to how many you got right, and Recall measures how many you got right that was actually right. AUC is a model that measures the area under an ROC curve, which measures the True Positive over False Positive Rate.

E2. This model does a good job of predicting whether or not a customer churns with an accuracy of 83%, Precision of 74%, Recall of 68%, and AUC score of .80. This implies that the model does well with predicting the accuracy of whether a customer churns or not but doesn't do so good at precision and Recall because most of the data points are skewed towards no making accuracy seem higher than it is. The AUC score shows the model quite a bit better than random chance, but it is still quite a way from being perfect.

A black screen with white text

Description automatically generated

A chart with numbers and labels

Description automatically generated with medium confidence

A graph of a curve

Description automatically generated

E3. A couple of limitations of this dataset include the fact that there might be better models. It only picks true values right .68%of the time, so there might be another model that, like Random Forest or Extreme Gradient boost, might be a better fit for this particular question. I also only cross-validated the data once, and this might be a split set that either increases or decreases model accuracy. It is always good to do multiple cross-validation with the same dataset to see if this model is not an outlier. Also, looking at the training and test set, the model performs close to 10% on most metrics, which might show signs of overfitting.

E4. The company should look at other models to predict or edit the hyperparameters or even add some other columns to see if you can get a more accurate model for this dataset. The model performs OK, is good enough to implement, and will likely get you good results. The model for this analysis can be improved by making a few adjustments.