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  
I will use the k-nearest neighbor classification method to help predict which customers are more likely to churn away and answer the question, what variables influence customer churn the greatest?

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.  
The goals of the data analysis are to determine which customers are most likely to churn away and looking at the commonalities between them. This will allow us to focus on the variables that influence customer churn the greatest so that we can try to reduce it.

**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 classification method allows the dataset to be analyzed by predicting the class labels for new datapoints based on how similar they are to their closest “k” datapoint. It does this by storing the entire dataset and, when a new data point needs to be classified, it calculates the distance between this query point and all data points in the training set. It then identifies the k closest data points and assigns the class label based on a majority vote among these neighbors. For this particular analysis I expect the algorithm to label customers that have higher monthly costs & higher GBs that are used per year as the ones that are more likely to churn away.

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

One assumption of the k-nearest neighbor classification method is that data points that are near each other are likely to be in the same classification class

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

Pandas: used for data manipulation and analysis

Numpy: used for numerical computations and handling large arrays or matrices

Sklearn: used for data preprocessing, model training, evaluation, standardization of numerical variables (StandardScaler), etc

Seaborn: used to create various, static, informative, etc, visualizations

Scipy: used for advanced mathematical and numerical operations

Pylab: graphs and plots

Statsmodels: for more statistical based models

**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.

Splitting the groups into training and test sets. The training set is used to teach the machine learning model and the test set is kept separate to evaluate the models performance.

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 numeric or categorical.

DummyChurn - Categorical

Children - Numeric

Bandwidth\_GB\_Year - Numeric

Age - Numeric

Tenure - Numeric

MonthlyCharge - Numeric

3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

\*The code is shown in attached notebook\*

First I performed some data cleaning processes (renaming columns, converting categorical variables, checking for null values, dropping unnecessary columns) to ensure that the data is free from errors, inconsistencies, and missing values. This makes sure that we have accurate model predicitons. Throughout this process I used the head function to make sure things were looking correct.

Next, since I am using the KNN classification method, I needed to standardize my data. First I initialized the standard scaler function. I then fit it onto the data to transform it. This changed the data into a numpy array. Since it did that I had to convert it back into a dataframe so I did that using pd.DataFrame().

Now that it was back into a data frame, I had to add the column names back into it. From here I used the head function to check on how everything looked and noticed that the Dummy Churn variable had decimals. I needed to convert them back to classes so I used pd.cut() to divide it back into classes (yes = 1 and no = 0).

4. Provide a copy of the cleaned data set.

\*Shown in attached notebook\*

**Part IV: Analysis**

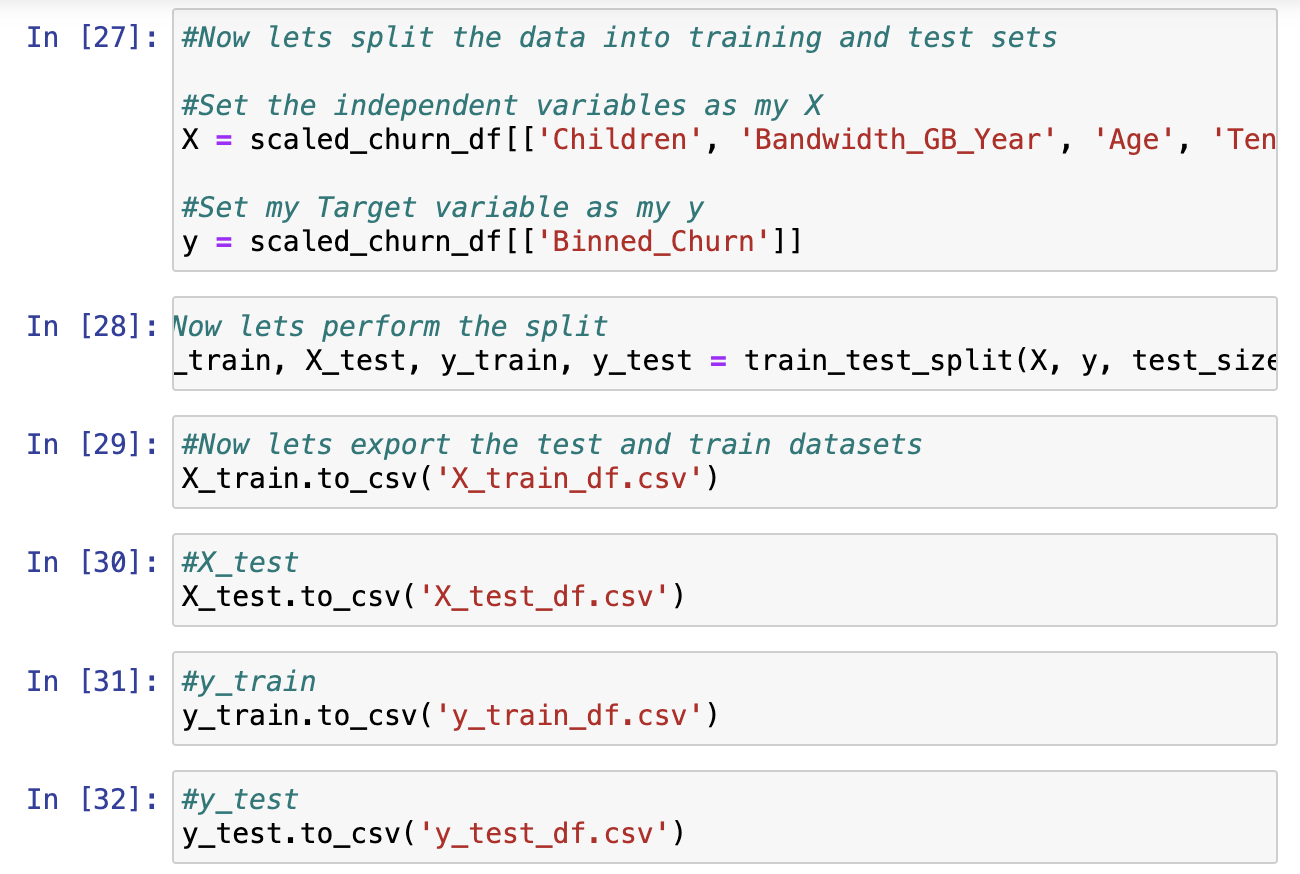
D. Perform the data analysis and report on the results by doing the following:

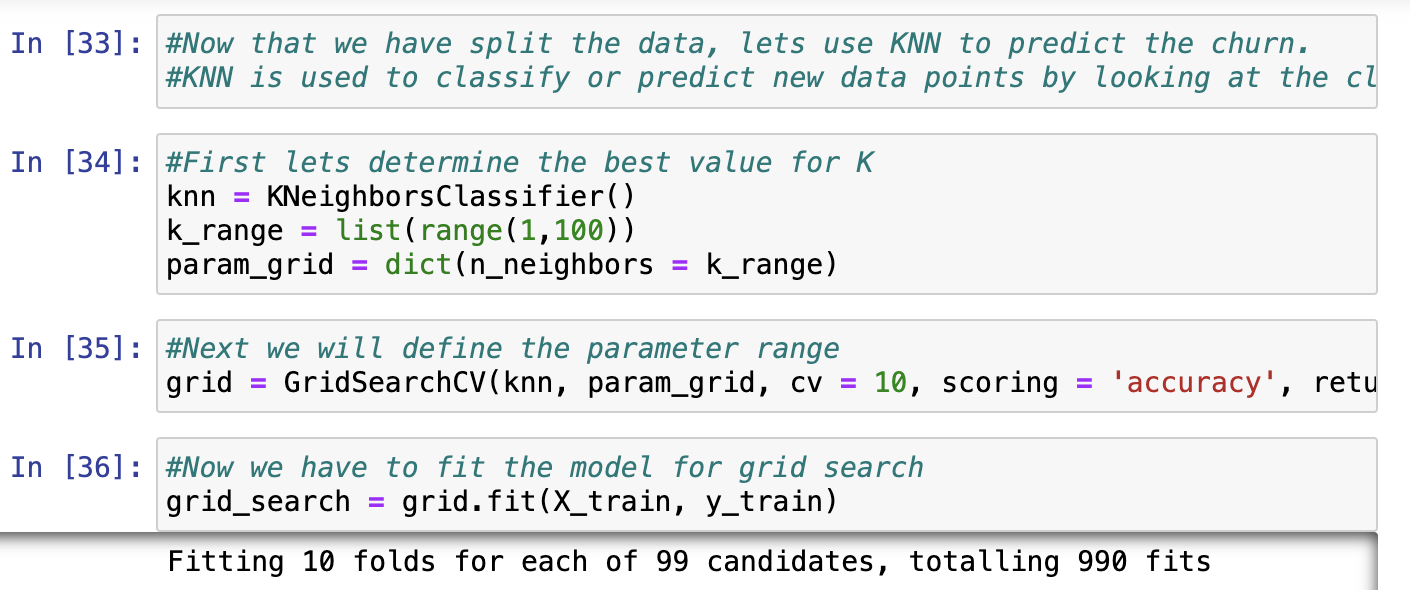
1. Split the data into training and test data sets and provide the file(s).

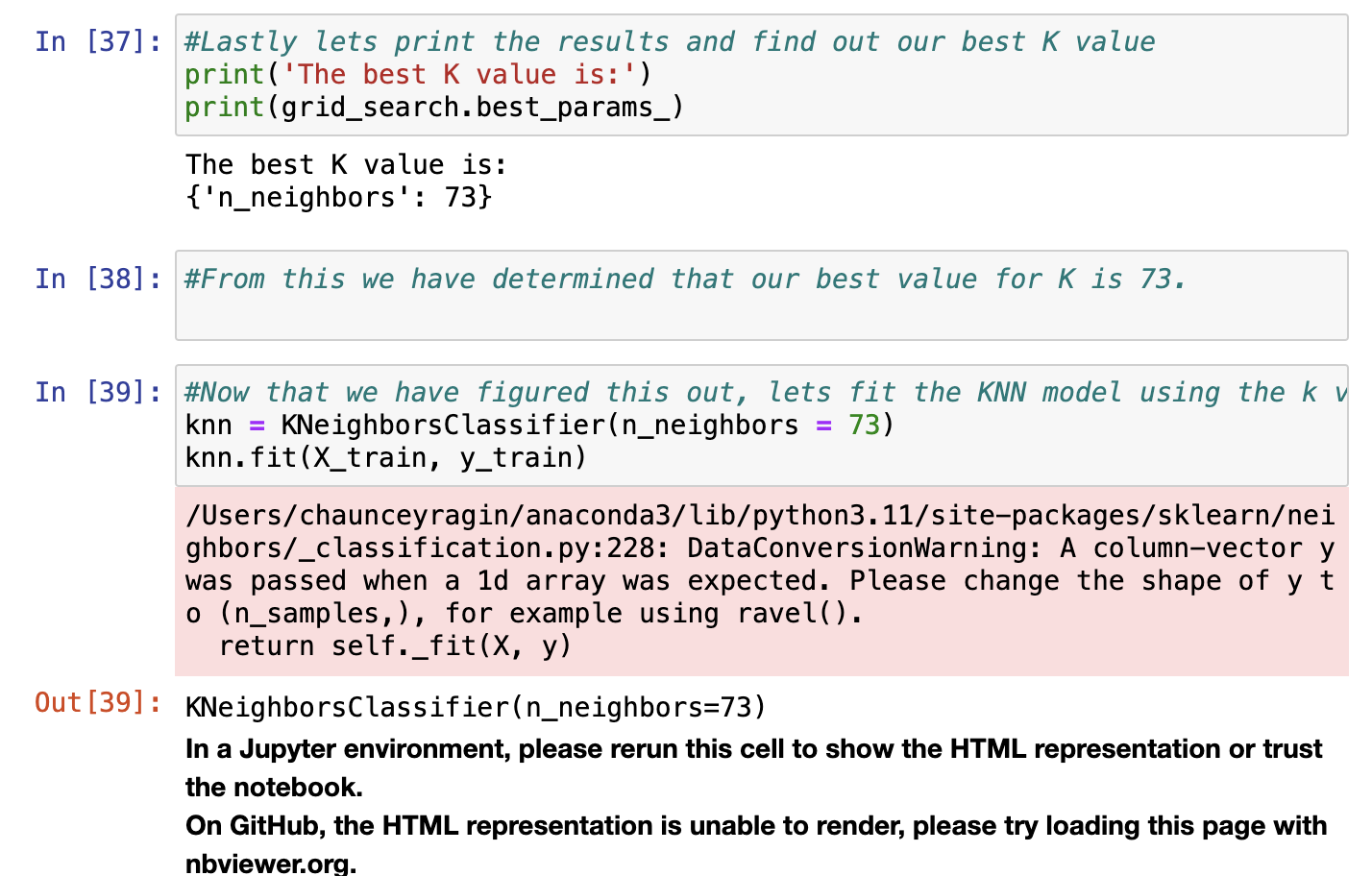
\*Shown in attached notebook\*

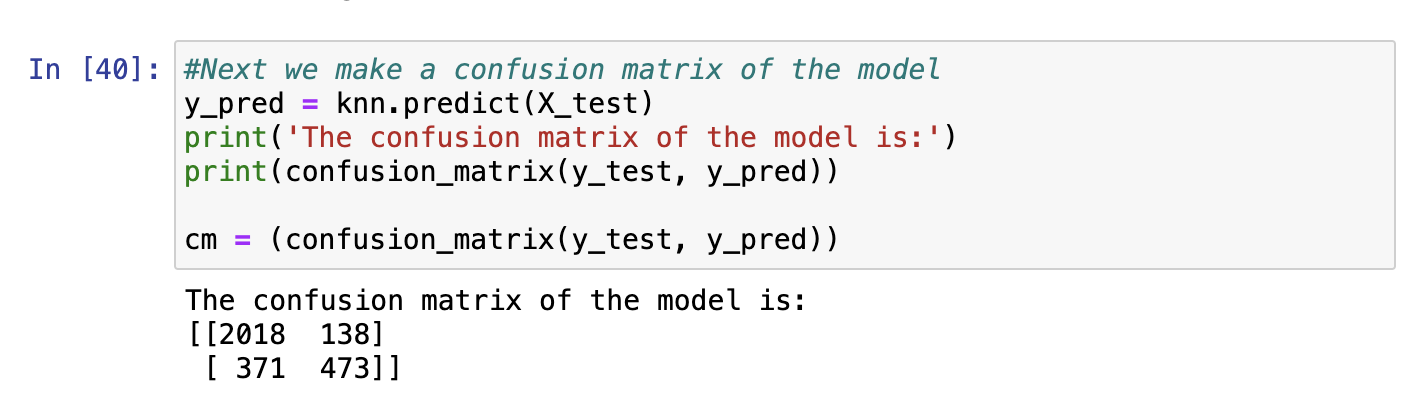
2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

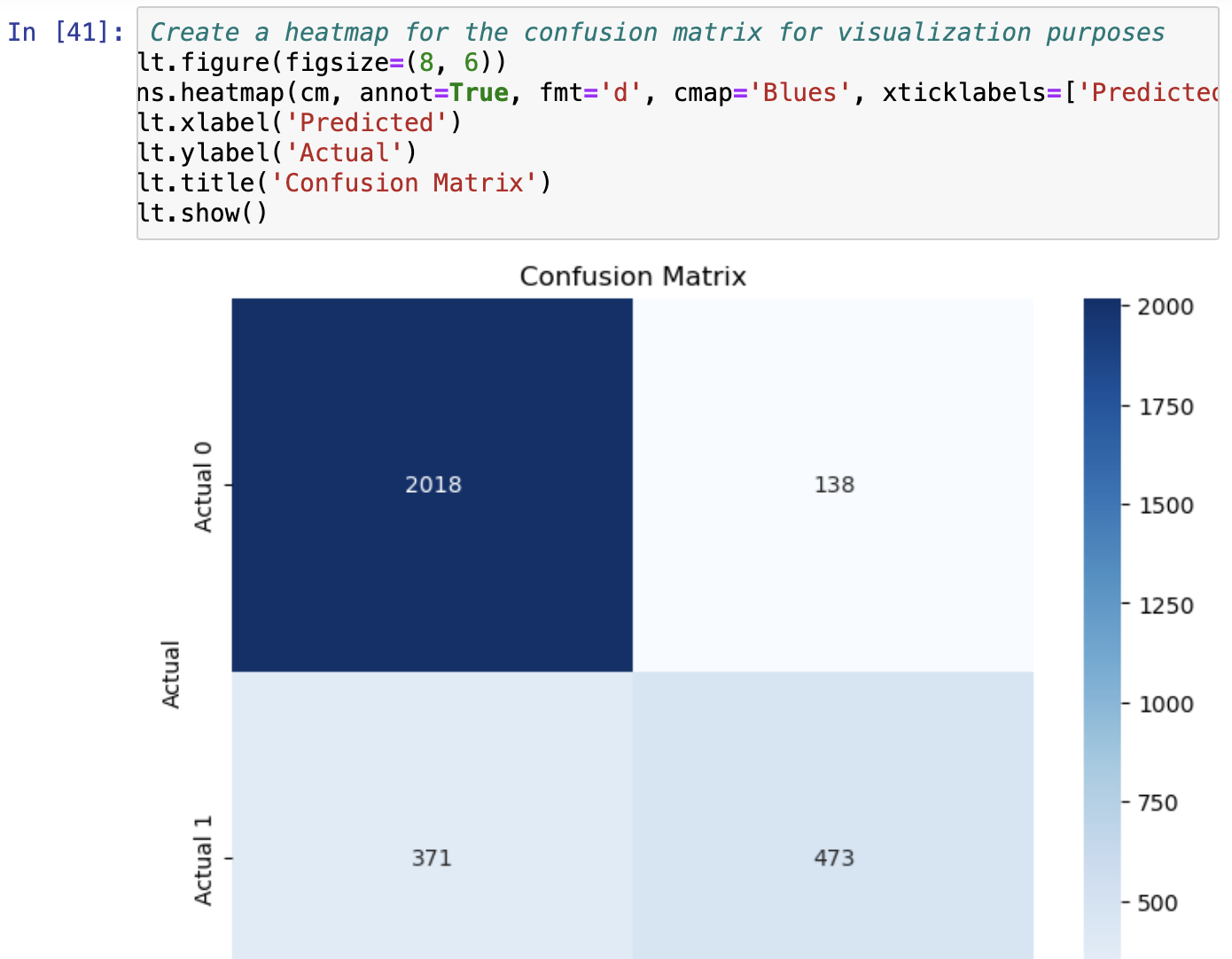
KNN is used to classify or predict new data points by looking at the closest data points in the feature space. It relies heavily on the choice of distance metric and the parameter k to appropriately analyze the data. See the below screenshots to see what I did:

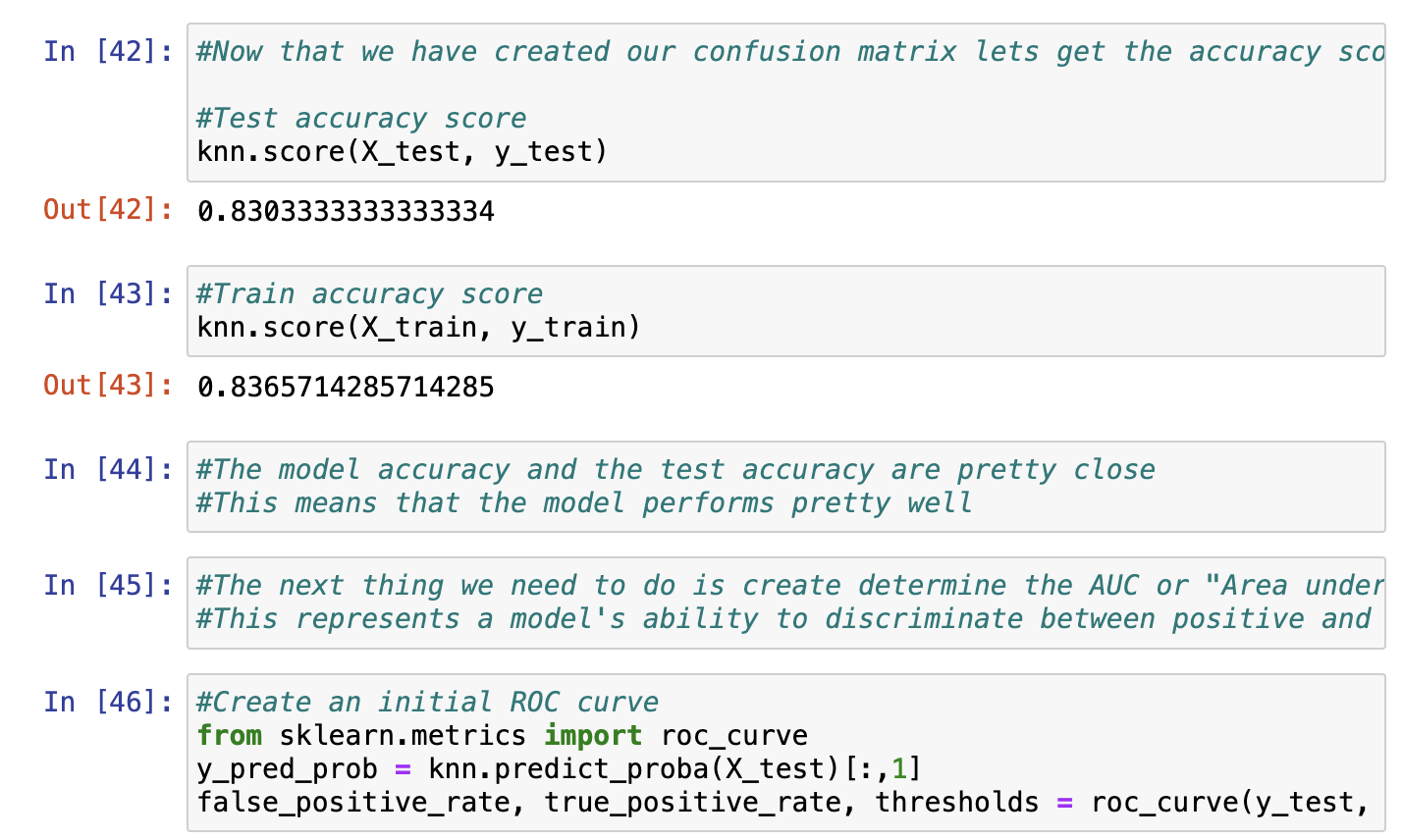


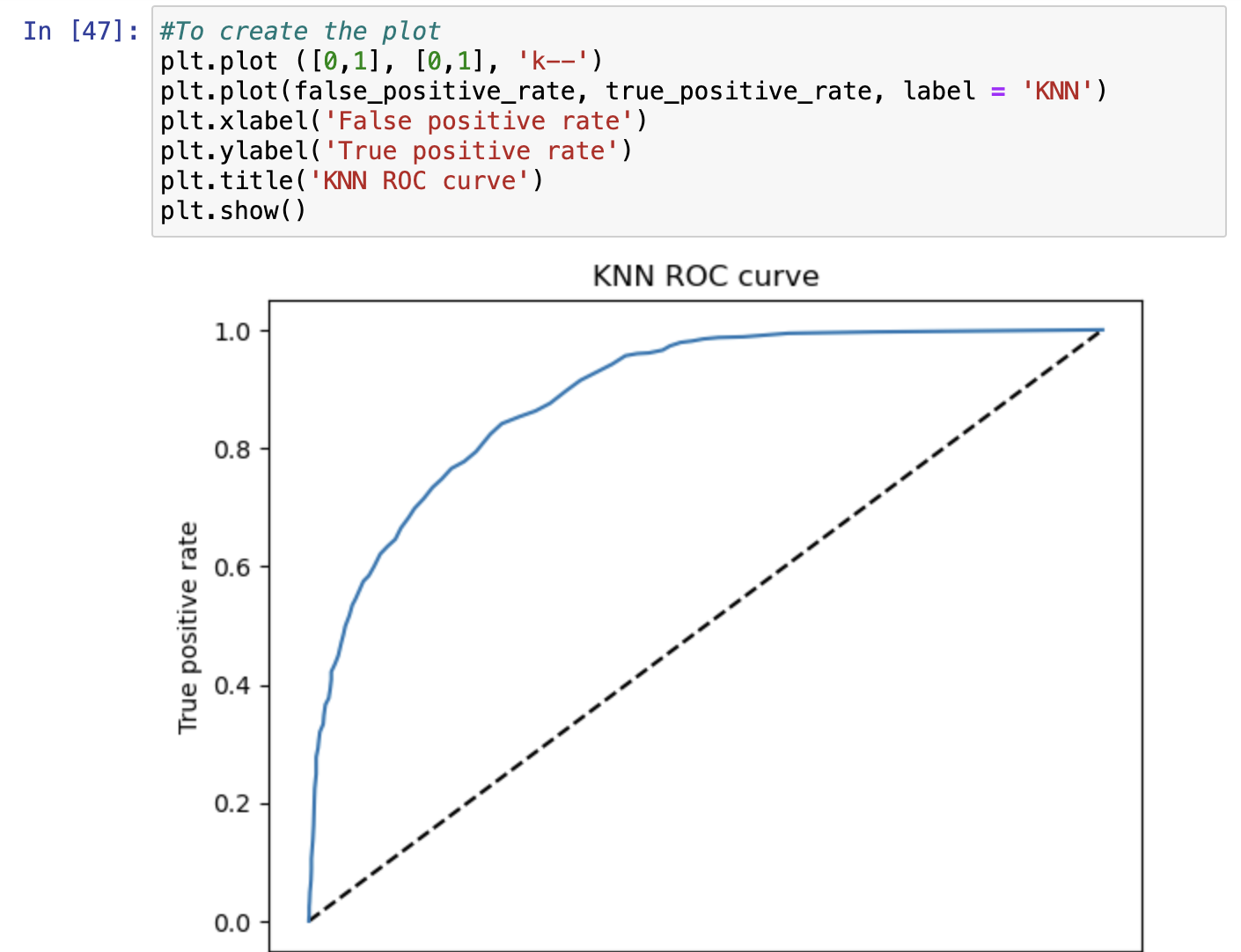


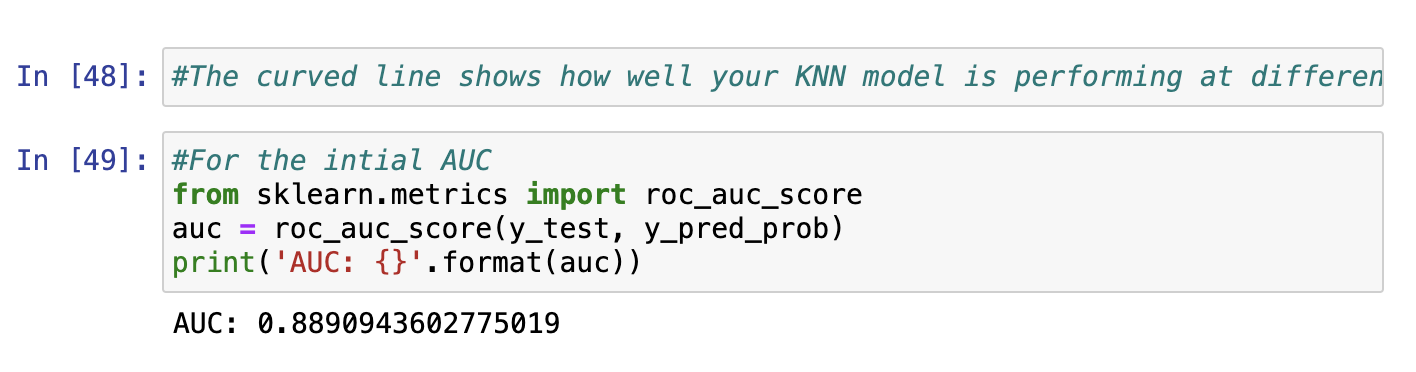












3. Provide the code used to perform the classification analysis from part D2.  
\*Shown in attached Notebook\*

**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 is the true positive plus the true negative which is then divided by the total number of instances. The AUC comes from the ROC curve. This curve shown in the plots is the true positive rate vs the true negative rate.

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

The AUC score I got came out to be .88 which is considered excellent. This means that our model is good at discriminating between positive and negative values.

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

One limitation that arises when using this method is that it can degrade in high dimensional spaces, meaning that it can struggle to find meaningful neighbors when this is the case leading to reduced accuracy values.

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.  
Using the KNN model that we created, a real world situation we could use this in would be to identify customers that were at risk and look for correlations in some of the other independent variables between them. For example, if the customers that are churning away all have a higher monthly payment amount, that would suggest that that value being higher is a good indicator of customer churn.

**Part VI: Demonstration**

F. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

*Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.*

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “*[*Data Mining I - NVMx | D209 (student creators) [assignments]*](https://wgu.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx#folderQuery=%22d209%22&folderID=%22dcfdf6dd-9e05-4665-b1b6-ac72018a20ab%22)*.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

G. Acknowledge web sources, using in-text citations and references, for segments of third-party code or data used to support the analysis. Be sure the web sources are reliable.

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I. Demonstrate professional communication in the content and presentation of your submission.