Answers of the Questions

1. Did you preprocess any of the features? If so, why? If not, why not?

Answer: Firstly, I fill the missing values of columns “total\_day\_charge” and “total\_eve\_calls” with mean values of the corresponding columns, then I do the one-hot encoding for the columns “state”, “area\_code”, “international\_plan” and “voice\_mail\_plan”. Because it the values are missing or the values are not numbers, the values cannot be input into the machine learning models (I use naïve bayes model to be the model).

2. Which features are the most relevant for predicting the output? How did you measure

feature importance?

Answer: We can see that from the code that the most relevant for predicting the output are “total\_eve\_minutes”, “total\_day\_minutes” and “total\_night\_minutes”. I measure the feature importance by the coefficient which is the feature log probability for interpreting MultinomalNB function in sklearn.

3. What metrics did you use to measure the performance of your model? How did you

determine how well your model generalizes?

Answer: I can use confusion matrix to measure my model within the measures like recall, precision, accuracy and so on. I can also draw and ROC figure and use AUC to measure model performance under difference criteria.

Note: I put all of the files in this file folder and compress it into a package I sent to you. However, your requirement does not want me to upload churn\_dataset\_train.csv. You can decompress it, add churn\_dataset\_train.csv to the same directory and directly run the code without the worry of directory.