Nabil Koney-Laryea, 13247508
CIS4930 Individual Coding Assignment
Spring 2023

1. Problem Statement

a. For the Python Fundamentals task we want to know more about the current state of the research topic “How have VR (virtual reality)/AR (augmented reality) techniques been used for education?” We have collected data from the past 5 years and need to analyze and visualize the data. This problem is important to solve if we want to publish in this field. I solved this problem by plotting figures to provide a better insight into where the publications are coming from. These figures also show which researchers, countries, and organizations are influential in the field.

b. For the Regression task we are conducting a study on the system usability of Siri to make flight recommendations. We want to know which features of the dataset are influential in predicting the usability rating a user gives to Siri. This is important to know so that future assistants such as Siri can improve with these features. I solved this problem by creating a linear regression model that can predict a SUS score based on the other variables in the data. This model can also report on the relationship between the independent variables and the dependent variable.

c. For the Classification task we take the same dataset, but this time we want to predict whether or not the user will purchase a plane ticket based on how well Siri understands human language, how well Siri understands the user’s intent, how long the user uses Siri for, and what the gender of the user was. This is an important problem because it could inform us on whether or not assistants such as Siri are obstacles to our goals. I solved this problem by implementing multiple classification models and evaluating which performed the best on the dataset.

2. Data Preparation

For each task I created a copy of the dataset. For each copy I dropped the columns that were irrelevant. For example, for the yearly publications visualization, I only selected the Year and Article No. columns. In some instances, I eliminated columns that differentiated different rows of data. Thus, after dropping columns from the dataset copies, I often found myself dropping duplicate rows as well.

Additionally, I removed rows that had missing data and I occasionally was required to reformat the data. Formatting country names from the article info dataset as country codes is one example of a time I had to reformat the data. I also reformatted the data for the regression and classification tasks by using ski-kit learn’s StandardScaler() to ensure all independent variables were within the same rang.

3. Model Development

- Model Training

For each model I created a training dataset and a testing dataset (test_size = .3) so that I could train the models and then test them on data they have never seen.

a. For the linear regression task on the siriTicketPurchases data I trained 2 different models. The first model was made using the sklearn linear regression model. The second model was built using the statsmodels library. I used the latter model for my evaluation due to it's easy to read summary.
b. For the classification task I used sklearn to create a Logistic Regression, SVM, Naive Bayes, and Random Forest model. The only difference between the training procedure for the classification models is that I scaled the data for the independent variables prior to creating the test/train split. I did this to ensure that the weights for the classification aren't skewed in the favor of variables with larger ranges.

- Model Evaluation
  a. To evaluate the linear regression models I looked at the R^2 scores. The statsmodels linear regression model scored an R^2 of .593. I believe a higher R^2 score would indicate good performance for this model. I did not experiment further to improve the model's performance.
  b. To evaluate the classification models, I first created a visualization of how many data points represented each class. I found the data to be fairly even and used this insight to decide on evaluating the models using the AUC for their ROC curves. The models performed very similarly to each other with values greater than .8 for precision and recall for both classes. Despite their similar performance, the Naive Bayes Classifier was shown to have a slightly higher AUC (.93) than the other models.

4. Discussion

a. I don't believe my model performs well enough. I believe that it wouldn't be accurate enough to predict the SUS score a user would give to Siri based on the independent variables chosen. In the notes of the OLS summary it was stated "[2] The condition number is large, 1.27e+03. This might indicate that there are strong multicollinearity or other numerical problems." This likely means that there needed to be additional steps taken to prepare the data.
b. I believe that the Naive Bayes classifier, as well as the other classifiers, performs very well on the data having a high precision, recall, F1 score, and AUC for the ROC curve.
c. During the data preparation process the greatest challenges faced were in the Python Fundamentals section. To create accurate visualizations the data had to be carefully merged, parsed, and cleaned. I had the most trouble creating the world map visualization as for any implementation there were several steps to geocoding and cleaning the Country data. To ensure my success, I double checked the counts of data for instances when I would have to remove duplicates or group the data to

minimize loss of entries. For visualizing the world map I relied on a country converter library to aid me in formatting the data.

d. During the model development process the main issue I ran into was evaluating the models. For both the regression and classification tasks, I was unsure of what metrics were best for what scenario. This was specifically an issue for understanding the OLS summary for the linear regression model. I mainly relied on online articles and previous statistics knowledge to effectively compare and assess the models I created.

5. Appendix

https://github.com/nabilkoneylaryea/mml-project-1