Gailyn Gabriel

Data Science Intensive Program

Predicting User demographics from phone behavioral data

1) With my project I am looking to predict user demographics from the way they engage with their phone apps.

2) My solution is valuable because it offers unique insight for marketers to build engaging and effective campaigns. With a deeper understanding of the user, not only will ad revenue have the potential increase, but may lead to greater satisfaction from the viewer due to a sense of personalization in content. Additionally, this model may offer constructive insight for developers and product design teams to create engaging products that resonate with specific users and encourage app usage and growth.

3) My data comes from kaggle by way of Talkingdata, China's largest third-party mobile data platform. Talkingdata holds behavioral data from over 500 million users . I will access the data using the Kaggle API.

4) To accomplish my task I will begin by cleaning and arranging the data. After unzipping the file, there are a number of different csv files containing all of the data required for the task.
Next, I will move into cleaning and merging the data. I will be checking for missing values, duplicate values, outliers, and decided how to handle them based on the distributions and occurrences within the dataset. Additionally, I will change data types to make visualization,analysis, and merging the csv files easier. I will use seaborn and matplotlib to visualize trends within the data. Furthermore, I will supplement my visualization with descriptive statistics about the dataset. This process may also highlight potential class imbalances that I will address. Next, I will create features to support my models, and will tweak according to my models performance. I will use the supervised models : Random Forest, Naive Bayes, Gradient Boosted, SVM, Decision Tree, Logistic Regression and KNN. I will use GridSearchCV for each model to find the ideal hyperparameters. I will use a 20% holdout group I plan to cross-validate with at least five folds. I will test these models for performance and evaluate the best performance using a classification report. Precision, F-1 score, and Accuracy will be important in distinguishing the best performing model. After initial testing, I may try dimensionality reduction methods to boost performance or decrease runtime. I will test two different methods: select K best and PCA. Additionally, I plan to use my deep learning specialization to boost performance. I will use a convolutional neural net with at least 10 epochs and evaluate based on score. Secondly, I will use unsupervised learning methods to further evaluate my data. I will use a clustering k means, meanshift, spectral clustering, and affinity propagation. I will evaluate the performance of these cluster methods based on similarity via silhouette score.

5) One of the biggest challenges I foresee will be managing runtimes due to the size of the dataset. Furthermore, an additional challenge will be in feature engineering if the performance of my models will not meet ideal standards. Hopefully PCA or selectk best will help remedy the problem.