UCLA Extension Data Science Intensive

Instructor: William Yu

Project 6

Use Python/Jupyter to do the following and submit your results in one Jupyter html file via Canvas with some explanations.

A. What Predicts the Long-term Home Price Appreciation of a City?

- Read my article: "R02_What Predicts the Long-term Home Price Appreciation of a City?" in Week 2's reading folder to understand the context of the data.
- In Script & Data 2, download W02b_homeprice.xlsx and save it into your computer.
- Write a Jupyter script to do the following:
- Plot a correlation chart for all the variables.
- Replicate Figures 1, and 5 in the article.
- Replicate a linear regression in Table 2.
- Make in-sample prediction of the linear regression model and check with "matrix_multiplication" tab in the W02b data file.

B. Logistic Regression: Churn Analysis

- Any successful businesses need to have a good customer service, management and analysis. How to **get** customers? How to **keep** customers? How to **grow** customers? The churn analysis is part of solutions to the question: how to keep customers?
- Revisit W04a_churn.csv. You can see the variable description here: https://www.kaggle.com/blastchar/telco-customer-churn
- The dependent variable (y) is churn: whether the customer left within last month.
- Write a Jupyter script to do the following:
- (1) Do some visualization and EDA for the variables.
- (2) Use sm.Logit library to run some logistic regressions (by including different combination of variables) in order to understand what variables are statistically significant predictors to predict churn. Note that in this project you don't need to divide the sample into trainset/testset. Show two best models you got. Use these two models to do the following:
- (3) Show the in-sample predicted probability of churn for each sample. And with the default threshold: 0.5 to show the predicted binary result.
- (4) Show the Confusion Matrix and ROC Curve and Accuracy, precision, recall and F1. Note: You should have two confusion matrix and two ROC curves.
- (5) With these two models, find the optimizing threshold based on F1 score. Use those new thresholds to redo prediction and show the recall and F1 scores.

C. Machine Learning Model -- Classification

- In j03_data script, we took a look at the Titanic data.
- In this project, we will revisit the data and run a bunch of machine models to predict who will survive in this historically tragic event.
- Follow j06_classifiction script, run Logit NBayes, LDA, QDA, KNN, and SVM model with a random selection of the train-set of 70% of the whole sample (30% test-set) and show kappa and accuracy of the test-set.
- Note that there are various parameter setting you can play with some machine learning models (called ***fine tuning***). Use Google search to see what kinds of model setting you can change and see how those changes impact your results.
- Select a best machine learning mode with the highest Kappa.
- Use that best model to do another train test split (with a different random_state) and show its accuracy score, confusion matrix, and classification report.