CASE STUDY # 1 (Group)

BUS 4023- WINTER 2021

15 Marks

DELIVERABLES FOR CASE STUDY INCLUDE

1) Answers in Markdown, Python code and output saved as notebook [in Jupyter notebook click on File > Download as > Notebook (.ipynb)]
2) Save the NOTEBOOK (.ipynb) file in the following format: **CASESTUDY_#_X_Y.IPYNB**
   Where,
   # -Case study number
   X=Group Number
   Y=Question Number
3) One Jupyter Notebook file for each question so total 2
4) One PowerPoint slide (for both questions) per group

Please remember to add the following on the first line of the notebook as a comment:

- your group name, names of all members
- the assignment problem number the notebook code and output pertain to

Please remember to add your group name and all member names to both the Power point presentation as well as the comment section of the Python code.

Case study covers materials from chapters 1,2,3,4 ,6 and 10
Topics covered: Descriptive statistics, Exploratory analysis and graphs, Variable reduction, Linear and logistic models, Performance Evaluation

CASE STUDY QUESTIONS (PAGE 527 OF TEXTBOOK)

21.3 TAYKO SOFTWARE CATALOGER

1) **7.5 Marks**
   - Develop a logistic regression model for classifying a customer as a purchaser or non-purchaser. Partition the data randomly into training set 60% validation set 40%. Run logistic regression with L2 penalty, using method LogisticRegressionCV. Please submit Python code. (5 Marks)
   - Tell a high-level story of steps taken to get to the end result. Start with the framework i.e., objective, exploration, variable selection (PCA, Correlation etc.). Then provide the final results and comparison analysis of the training vs. validation data vs. test.
   - Present your findings in power point format (no more than 5 slides) in terms of steps taken and results. (2.5 Marks)

Things you can add:

- Show the shape of the df
- Show some records of the df
- List data types of the variables in the df
- Preliminary Exploration - view the data: rename all column names - replace space with underscore
- Look at descriptive statistics
- Count of Missing values
- Remove certain variables from the onset **(i.e., spending and sequence number)**
- Count number of unique values in each variable
- Dummy variables if need
- Some visualizations to explore the data
    - Histograms, Frequency Distribution, side by side plots with the outcome, scatterplot, pairplot
    - Other plots according to your discretion
- Correlation table & Heatmap: Comment on high correlations
- Conduct a PCA: Discuss how many PCs to use

The Logistic Regression:

- (don't incorporate PCA or variable reduction through correlations here, run the logistic regression on all variables apart from spending and sequence_number)
- Partition the data on the whole data set randomly into a training set 60% validation set 40%
- Run quick descriptive stats for validation and training dataset
- Fit a logistic regression (set penalty=l2 and C=1e42 to avoid regularization): Predict the model on validation dataset
- Develop gains and lift chart for test and validation results
- Confusion matrix for all sets
- Show some use of statsmodel if possible

2) **(7.5 points)**
   - Develop a model for predicting spend among purchasers. Refer to problem #3 in case study. Create subsets of the training and validation sets for only purchasers' records by filtering for Purchase = 1. Develop models for predicting spending with the filtered datasets, using: Multiple linear regression (use stepwise regression). Choose one model on the basis of its performance on the validation data. Please submit Python code. (5 Marks)
   - Tell a high-level story of steps taken to get to the end result. Start with the framework i.e., objective, exploration, variable selection (PCA, Correlation, Exhaustive search etc.). Then provide the final results and comparison analysis of the training vs. validation data vs. test.
   - Present your findings in power point format (no more than 5 slides) in terms of steps taken and results (2.5 Marks)

***Decision Trees not required**

Things you can add:

- Show the shape of the df
- Show some records of the df
- List data types of the variables in the df
- Preliminary Exploration - view the data: rename all column names - replace space with underscore
- Look at descriptive statistics
- Count of Missing values, discuss briefly on this, what would you do if there are/were missing values
- Remove certain variables from the onset **(i.e., purchase and sequence number)**
- Count number of unique values in each variable
- Create dummy variables if need, discuss briefly on this
- Some visualizations to explore the data
    - Histograms, Frequency Distribution, Side by side Box plots with the outcome, Scatterplot, pairplot
    - Other plots according to your discretion
- Correlation table & Heatmap: Comment on high correlations
- Conduct a PCA: Discuss how many PCs to use

The Linear Regression:

- (Don't incorporate PCA or variable reduction through correlations here, run the linear regression on all variables apart from purchase and sequence_number)
- Partition the data on the whole data set randomly into a training set 60% validation set 40%
- Run quick descriptive stats for validation and training dataset
- Fit a linear regression: Predict the model on validation dataset
- Create histogram using residuals
- Choose the best model using forward selection,backward elimanation and stepwise selection method
- Show regression summary using the best model
- Develop a lifts and gains chart using the best model