MMAI 5040: Business Application of AI 1 (Winter 2022)

Assignment 1 (60 points) – Worth 15% of Final Grade

This assignment is to be done individually. Submit your answers in a either 1) a word document or pdf and include your Python Code as a Jupyter notebook, or 2) a single Jupyter notebook containing all your answers and python codes.

Name the file(s) as follows: "MMAI5040_W22_Student_Name_Assignment1"

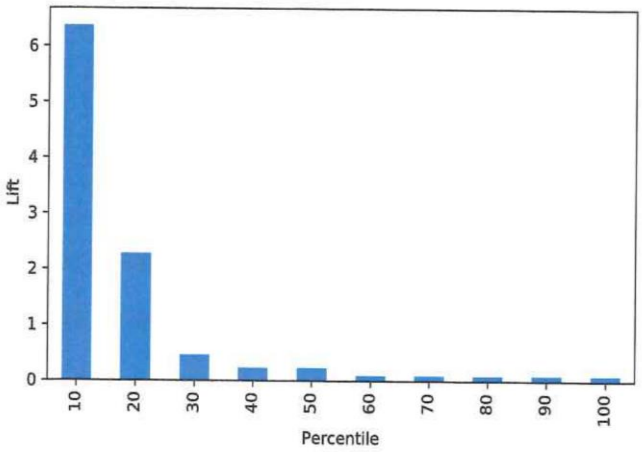
Grading will be based:

- Accuracy, clarity and precision of conceptual arguments and strategies.
- Correctness (rather than quality) of implementation in Python
- Originality in discussions of the business implication of outputs.
- Organization and presentation

Due Feb 14 @ 11.59pm. The usual penalty of 20% daily late submission applies.

Q1. **[15 points]** A machine learning algorithm has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly so) and 952 as non-fraudulent (920 correctly so).

a. Construct the confusion matrix and calculate the following metrics: overall error rate, sensitivity, specificity, and lift.
b. Suppose that this routine has an adjustable cutoff (threshold) mechanism by which you can alter the proportion of records classified as fraudulent. Describe how moving the cutoff up or down would affect
    i. the classification error rate for records that are truly fraudulent
    ii. the classification error rate for records that are truly nonfraudulent
c. Below is the decile lift chart for applying the model to classify new data.



    i. Interpret the meaning of the first and second bars from the left.

ii. Explain how you might use this information in practice.
iii. Another analyst comments that you could improve the accuracy of the model by classifying everything as nonfraudulent. If you do that, what is the error rate?
iv. Comment on the usefulness, in this situation, of these two metrics of model performance (error rate and lift).

Q2. [20 points] Predicting Airfare on New Routes

The following problem takes place in the United States in the late 1990s, when many major US cities were facing issues with airport congestion, partly because of the 1978 deregulation of airlines. Both fares and routes were freed from regulation, and low-fare carriers such as Southwest (SW) began competing on existing routes and starting nonstop service on routes that previously lacked it. Building completely new airports is generally not feasible, but sometimes decommissioned military bases or smaller municipal airports can be reconfigured as regional or larger commercial airports. There are numerous players and interests involved in the issue (airlines, city, state and federal authorities, civic groups, the military, airport operators), and an aviation consulting firm is seeking advisory contracts with these players. The firm needs predictive models to support its consulting service. One thing the firm might want to be able to predict is fares, in the event a new airport is brought into service. The firm starts with the *Airfares.csv* dataset, which contains real data that were collected between Q3—1996 and Q2—1997. The variables in these data are listed in Table 2 and are believed to be important in predicting FARE. One question that will be of interest in the analysis is the effect that the presence or absence of Southwest has on FARE.

a. Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?
b. Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE?
c. Find a model for predicting the average fare on a new route:
    i. Convert categorical variables (e.g., SW) into dummy variables. Then, partition the data into training and validation (40%) sets. The model will be fit to the training data and evaluated on the validation set.
    ii. Use stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S_CODE, S_CITY, E_CODE, E_CITY). Report the estimated model selected.
    iii. Repeat (ii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (ii) in terms of the predictors that are in the model.
    iv. Compare the predictive accuracy of both models (ii) and (iii) using measures such as RMSE, average error and lift charts.
    v. Using model (iii), predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = 28,760, E_INCOME = 27,664, S_POP = 4,557,004,

E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

vi. Predict the reduction in average fare on the route in (v) if Southwest decides to cover this route [using model (iii)].

vii. Realistically, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?

viii. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search to find such a model.

ix. Use the model in (viii) to predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = 28,760, E_INCOME = 27,664, S_ POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12782, DISTANCE = 1976 miles.

x. Compare the predictive accuracy of this model with model (iii). Is this model good enough or should it be re-evaluated once flights begin on the new route?

Table 2. Description of variables in the Air Fare dataset.

S_CODE	Starting airport's code
S_CITY	Starting city
E_CODE	Ending airport's code
E_CITY	Ending city
COUPON	Average number of coupons for that code (a one-coupon flight is a nonstop flight, a two coupon-flight is a one-stop flight, etc.)
NEW	Number of new carriers entering that route between Q3—1996 and Q2—1997
VACATION	Whether (Yes) or not (No) a vacation route
SW	Whether (Yes) or not (No) Southwest Airlines serves that route
HI	Herfindahl index: meaure of market concentration
S_INCOME	Startying city's average personal income
E_INCOME	Ending city's average personal income
S_POP	Starting city's population
E_POP	Ending city's population
SLOT	Whether or not either endpoint airport is slot-controlled (this is a meaure of airport congestion)
GATE	Wether or not either endpoint airport has gate constraints (this is another measure of airport congestion)
DISTANCE	Distance between two endpoint airports in miles
PAX	Number of passengers on that route during period of data collection
FARE	Average fare on that route

Q3. [10 points] Identifying Course Combinations and Recommending Courses

The Institute of Statistical Education at Statistics.com offers online courses in statistics and analytics and is seeking information that will help in packaging and sequencing courses. Consider the data in the file *Courstopics.csv*. These data are for purchases of online statistics courses at Statistics.com. Each row represents the courses attended by a single customer.

a. The firm wishes to assess alternative sequencings and bundling of courses. Use association rules to analyze these data and interpret several of the resulting rules (i.e., as many as make sense to you).
b. The firm now wants to provide a course recommendation to a student who purchased the Regression and Forecast courses. Apply user-based collaborative filtering to the data. All recommendations will be 1. Explain why this happens.

Q4. [15 points] Pharmacy Industry

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures. Financial data gathered on 21 firms in the industry are available in the file *Pharmaceuticals.csv*. For each firm, the following variables are recorded:

1. Market capitalization (in billions of dollars)
2. Beta
3. Price/earnings ratio
4. Return on equity
5. Return on assets
6. Asset turnover
7. Leverage
8. Estimated revenue growth
9. Net profit margin
10. Median recommendation (across major brokers)
11. Location of firm's headquarters
12. Stock exchange on which the firm is listed

Use cluster analysis to explore and analyze the given dataset as follows:

a. Use only the numerical variables (1 – 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.
b. Interpret the clusters with respect to the numerical variables used in forming the clusters.
c. Is there a pattern in the clusters with respect to the categorical variables (10 – 12)? (i.e., those not used in forming the clusters).
d. Provide an appropriate name for each cluster using any of or all the variables in the dataset.