# BUS 41204 Machine Learning

Midterm

- **This is an INDIVIDUAL exam. You cannot work in groups**.
- The exam must be submitted on gradescope before 11.59pm on Sunday, February 23. You should submit a pdf document.
- Ask coding questions on Piazza. Do not reveal answers when formulating questions.
- When answering questions, provide plots and supporting analysis. Label plots and axes.
- Be concise.
- Provide supporting code to help us understand your answers.

# 1 Question [30 points]

Universal Bank has begun a program to encourage its existing customers to borrow via a consumer loan program. The bank has promoted the loan to `5000` customers, of whom `480` accepted the offer. The data are available in file `UniversalBank.csv`. The bank now wants to develop a model to predict which customers have the greatest probability of accepting the loan, to reduce promotion costs and send the offer only to a subset of its customers.

You will develop several models, then combine them in an ensemble. Partition the data into 60% training and 40% validation. Ignore the zip code column when building models. Build the following three models:

- logistic regression;
- `k`-nearest neighbors;
- classification trees.

Use `Personal Loan` as the outcome variable. Use all variables to build models, that is, you do not need to select variables. Do not spend much effort in finding the "best" tuning parameters for these methods.

Report the validation confusion matrix for each of the three models. Use a default threshold of 0.5 for converting predicted probabilities to binary outcomes.

Create a data frame with the actual outcome, predicted outcome and probabilities estimated by each of the three models.

Add two columns to this data frame for (1) a majority vote of predicted outcomes, and (2) the average of the predicted probabilities. Using the classifications generated by these two methods derive a confusion matrix for each method and report the overall accuracy. Report the first 10 rows of this data frame.

Compare the error rates for the three individual methods and the two ensemble methods.

# 2 Question [30 points]

Your company has hired a consulting firm to help with the fraud problem. You are present in the meeting where the consulting firm presents the results of their pilot study, showing that the model has a very low error rate (percent incorrectly classified instances). They argue to your boss that based on this great performance, she should hire them to build the system. You need to explain to your boss the notions of false positive and false negative errors, and how the system should be evaluated. You may assume that the only relevant decision is (binary): if the system predicts fraud, block the account; if the system predicts no fraud, do nothing.

Explain the following notions:

a) describe the confusion matrix to your boss;
b) describe how you will fill out the confusion matrix for the consultant's model;
c) describe the cost/benefit matrix for this problem;
d) explain briefly why the confusion matrix and the cost/benefit matrix are important for this problem (1-2 sentences);
e) show the proper evaluation function (equation) for the consultant's model;
f) how do the confusion and cost matrices come into play in this function.

# 3 Question [30 points]

## 3.1 Data

`EastWestAirlines.csv` is the dataset for this question. It contains information on 3999 passengers who belong to an airline's frequent flier program. For each passenger, the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.

Description of variables is provided in the table below.

| Variable Name | Data Type | Description |
|---|---|---|
| ID# | NUMBER | Unique ID |
| Balance | NUMBER | Number of miles eligible for award travel |
| Qual_miles | NUMBER | Number of miles counted as qualifying for Topflight status |
| cc1_miles | CHAR | Number of miles earned with freq. flyer credit card in the past 12 months |
| cc2_miles | CHAR | Number of miles earned with Rewards credit card in the past 12 months |
| cc3_miles | CHAR | Number of miles earned with Small Business credit card in the past 12 months |
| Bonus_miles | NUMBER | Number of miles earned from non-flight bonus transactions in the past 12 months |
| Bonus_trans | NUMBER | Number of non-flight bonus transactions in the past 12 months |
| Flight_miles_12mo | NUMBER | Number of flight miles in the past 12 months |
| Flight_trans_12 | NUMBER | Number of flight transactions in the past 12 months |
| Days_since_enroll | NUMBER | Number of days since Enroll_date |
| Award? | NUMBER | Dummy variable for Last_award (1=not null, 0=null) |

*Note*: variables `cc1_miles`, `cc2_miles`, and `cc3_miles` are binned.

```
miles bins: 1 = under 5,000
            2 = 5,000 - 10,000
            3 = 10,001 - 25,000
            4 = 25,001 - 50,000
            5 = over 50,000
```

## 3.2 Assignment

Use `k`-means clustering to identify clusters of passengeres.

- When choosing a good value for the numbers of clusters, `k`, think about how the clusters would be used. It is likely that the marketing efforts would support targeting only a few clusters.
- Consider scaling variables.

What would happen if the data were not normalized?

Which clusters would you target for offers, and what types of offers would you target to customers in that cluster?

# 4 Question [40 points]

You will use data in `marketing.csv` for this question. This dataset contains `64,000` customers who last purchased within twelve months. The customers were involved in an e-mail test: 1/3 were randomly chosen to receive an e-mail campaign featuring a discount offer; 1/3 were randomly chosen to receive an e-mail campaign featuring a buy one get one free offer; 1/3 were randomly chosen to not receive an e-mail campaign. During a period of two weeks following the e-mail campaign, results were tracked. The first 8 columns provide individual-level data and `conversion` column is the label that we will try to predict:

- `recency`: months since last purchase
- `history`: $value of the historical purchases
- `used_discount`: indicates if the customer used a discount before
- `used_bogo`: indicates if the customer used a buy one get one before
- `zip_code`: class of the zip code as Suburban/Urban/Rural
- `is_referral`: indicates if the customer was acquired from referral channel
- `channel`: channels that the customer is using, Phone/Web/Multichannel
- `offer`: the offers sent to the customers, Discount/But One Get One/No Offer

The data were collected through an experiment that allows us to find growth opportunities. Splitting the customers who we are going to send the offer into test (groups receiving one of the two offers) and control groups helps us to calculate incremental gains of offers.

a) Does giving an offer increase conversion? If yes, what kind of offer performs best? Discount or Buy One Get One?

For the rest of the question, we will focus on the customers who received the `Discount` offer or did not receive any offer.

b) Partition the data into 60% train and 40% validation set.

c) Build a model for $P(\text{conversion} = 1 \mid \text{offer}, x)$, where $x$ denotes the individual-level characteristics. Discuss how you chose the parameters used to build the model. Which one is it? Why did you choose it?

Our interest is not just in how different offers did overall, nor is it whether we can predict the probability that a customer will convert after receiving the offer. Rather our goal is to predict how much (positive) impact the offer will have on a specific customer. That way the marketing campaign can direct its limited resources towards the customers who are the most persuadable—those for whom sending the offer will have the greatest positive effect.

d) For each record in the validation set, compute the uplift defined as

$$\text{uplift}(x) = P(\text{conversion} = 1 \mid \text{offer} = \text{Discount}, x) - P(\text{conversion} = 1 \mid \text{offer} = \text{No Offer}, x).$$

If a campaign has the resources to target 25% of the customers, what uplift cutoff should be used? If we are to target 25% of the customers in the validation set based on the uplift obtained from your model, how much better would we do compared to random targeting?