UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

ASSIGNMENT 5

The goal of this assignment is to recap and review the methods we have covered, and provide you with an opportunity to reflect on and apply your learning in a hand-picked dataset. In this final assignment, you will approach an open-ended data analysis problem as it appears to you in the wild (as the analytics problems will often do so, evidently). Remember we have covered the essentials such as cross validation, resampling, ROC, etc. in addition to the following methods:

1. Linear regression
2. Logistic regression
3. Lasso regression
4. Ridge regression
5. Elastic net regression
6. Linear discriminant analysis
7. Quadratic discriminant analysis
8. k-nearest neighbors
9. Decision trees
10. Bagged trees
11. Random forests
12. Adaptive boosting
13. Gradient boosting
14. Time series regression
15. ARIMA models

Even without the ongoing **Text Mining** and **Neural Networks** added, this is quite a long list and exceptional ROI for a -disrupted- semester. You know how to build these models and how they work, underlying intuition, how to interpret (if interpretable), along with business use cases!

The dataset in this assignment is hand-picked (but **not** hand-crafted) to reinforce all of this learning in solving an interesting business problem using **real data**. As you work on the problem, you will see how much of a difference it makes to have a relatively balanced outcome variable.

Data dictionary

The data file is **airlines.csv** contains data from 3,999 airline customers enrolled in the program. Note that **the data is from a real airline reward program**, but the name of the airline is changed to East-West and the identifiers for customers are anonymized. The data contains information about each customer's history, as listed below:

Variable	**Description**
ID	Unique ID
Balance	Number of miles eligible for award travel
Qual_miles	Number of miles counted as qualifying for Topflight status
cc1_miles	Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles	Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles	Number of miles earned with Small Business credit card in the past 12 months:
Note: _miles are binned	*1 = under 5,000*
	2 = 5,000 - 10,000
	3 = 10,001 - 25,000
	4 = 25,001 - 50,000
	5 = over 50,000
Bonus_miles	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	Number of flight miles in the past 12 months
Flight_trans_12	Number of flight transactions in the past 12 months
Days_since_enroll	Number of days since Enroll_date
Award	Dummy variable for travel award claimed (1 = award claimed, 0 = not claimed)

East-West Airlines has two main goals: (1) identifying if a customer will claim a travel award using their rewards, and (2) identifying factors that lead to customers claiming a travel award.

Assignment Instructions

The objective of this assignment is to help East-West Airlines with their goals. This objective clearly involves running data analysis models with Award variable being the dependent variable. Beyond this, the assignment is open ended and requires you to think (and tinker) analytically.

Before you start: Load the libraries in the given order: *tidyverse, tidymodels, plotly, skimr, caret*

1) (0 points) Data preparation

a) Load the dataset into R and call it *df*. Inspect the data and do the preprocessing.
b) Set the seed to **123**.
c) Randomly split the dataset into a training dataset and a test dataset. Use **70%** of the data as training data and set the remaining **30%** as test data.

2) (5 points) Appropriate data analysis methods

Consider the problem of analyzing the East-West Airlines travel award data. Explore your data but **do not build any models yet**. You can check descriptive statistics, create box plots, scatter plots, frequency tables, and etc. Of all the ways we have used to **explore** data, choose three that you think would be appropriate for this problem, and explain for each:

a) Why do you think this exploration method is appropriate for the problem?

➔ *I used the following exploration plots for understanding the data,*

- ***Box Plots:*** *Box plots using balance miles, bonus from non-flight transactions, bonus from flight transactions. These Box plots tell us if there is a clear indication that these variables have an indication of how and when people redeem their miles. Based on these plots we can see that people who have redeemed the miles are people with more balance and more bonus gained over the period. So, we could say that people wait to grow their balance before they redeem it.*
- ***Scatter Plot:*** *The Scatter plot of "award" with "balance" on x-axis and "flight_miles_12mo" on y-axis shows us how people behave when they have some balance and how recently gaining some balance affects their decision. We can see that for people with lower balances a recent gain in miles during the past 12 months has a higher chance of redeeming the miles. While people with already higher balance are impacted in same way as they may or may not redeem the miles based on what their target is.*
- ***Histogram:*** *The Histogram with count of people redeeming and not redeeming based on how long ago they enrolled in the program tells us about the patterns. We see that percentage of people redeeming the miles are high for people who just joined and have been for more than 6000 days. The people in the middle are usually less inclined to spend miles as they are waiting for it to grow or trying to hit a target of miles before redeeming.*

b) How would you use this exploration method to address either Goal #1 of East-West Airlines, Goal #2 of East-West Airlines, or both?

➔ *The exploration methods listed above can address both the Goals. They tell us how the variables impact the DV. We can say that balance, time spent in the program and miles earned in past 12 months are indicators of the people who redeem miles. Also based on these values we could make a conclusion on whether a person would redeem their miles or not.*

3) (5 points) Goal #1: Identifying if a customer will claim a travel award using their rewards

a) If East-West Airlines is most concerned with identifying if a new customer will claim a travel award using their rewards, what three methods would you suggest as the most appropriate for this problem? Why?

➔ *We need to make predictions about whether a new customer will claim a travel award or not. We basically need a model with a high prediction accuracy. For a prediction model the variable significance and selection can be considered as the less significant part of the modelling. I picked the Logistic Regression, Random Forest and XGBOOST as my three models of choice. Logistic regression gives us the flexibility in case we do not want to use specific predictors or want to see the predictors impact on odds individually. While Random forest targets to improve the accuracy by trying many possible scenarios. Xgboost is a black box and we do not have much control but it does give us the variable importance data quickly.*
Other than these I also built the decision tree, and Knn model as predictive modeling is always a trial and error process to find the best model.

b) Run each of your three methods and decide which method East-West Airlines should use in practice. Provide a detailed summary comparing the information provided by the model summaries and performance metrics to support your answer.

➔ *On checking for the bias in the test data we see that the model has ~62% people who will not redeem and ~38% who would. So, for accuracy to be a relevant metric it has to be at least greater than 65% as 62% can be achieved by default. We will also check the sensitivity and specificity of the models. Based on the performance metrics of the models built we see that Xgboost gives us the highest accuracy. We can get the highest specificity by using Knn and highest sensitivity by using Random forests.*
Now for us, we should be able to identify the people who are willing to redeem the reward. So, we can be okay with some number of false positives, but we should not have false negatives. Thus, focusing on sensitivity we can choose the random forest model which gives us almost similar accuracy with higher sensitivity.

Hint: As we have discussed several times, paying attention to the variables you choose to include in a model is important. For example, if one wants to run a QDA model on this data,

cc3_miles must be dropped because the variable has the same value for 99.5% of all customers, leading to an imbalance QDA is not able to handle. Remember, the only assumption QDA makes is the normality of predictors given a dependent variable.

4) (5 points) Goal #2: Identifying factors that lead to customers claiming a travel award

a) If East-West Airlines is most concerned with identifying features that lead to customers claiming a travel award, what three methods would you suggest as the most appropriate for this problem? Why?

➔ *Here the goal is not predictive accuracy but identifying what factors impact the decision to redeem or not the most. For this the focus is shifted from the outcomes to the process and inputs. We need visibility on the inputs and how significant they are towards modelling. By picking the best variables with most importance we can build the most efficient explanatory models. The models I used are basic Linear probability model to get a feel of the collinearities in the data and variable significance based on p-values and coefficients. Another model I used is the lasso model which drops the insignificant variables by penalizing them. In our case we have factor variables which have multiple levels. On analyzing the lasso output, I saw that some of the levels were being dropped which makes explaining their significance difficult. Thus, I built another grouped lasso model to get the variable significance for factors as a whole.*
We could also used stepwise regressions, ridge or elastic net models for variable selection and significance. Also, the prediction models like logistic and xgboost can also give the variable importance but cannot explain their effects in magnitude.

b) Run each of your three methods and decide which features East-West Airlines should focus on for determining if customers will claim a travel award. Provide a detailed explanation to support your answer.

➔ *Based on the outputs from the group lasso model the features that East-West Airlines should focus on are, qual_miles, bonus_miles, bonus_trans, flight_trans_12, days_since_enroll. We can say that the balance is somewhat important as it has a low p-value in the Linear probability model but both lasso and group lasso show that its not that significant. Also, the factor variables cc1, cc2 and cc3 are not significant as a whole factor. This can be seen from both their p-values in linear probability model and in the group lasso outputs. Some of the levels are significant as per the output in the lasso model but we cannot explain the overall variable significance as other levels are dropped in this model*

Hint: Explanatory focus in the analysis and variable/feature importance are some ways to identify the most important variables/features for a given dependent variable and dataset.