**York University**

**Schulich School of Business**

**Course:** MMAI 5090 - Business Applications of Artificial Intelligence II

**Project Title:** Customer base segmentation by logistic regression

**Submission Date:** 28th July 2021

**Group 7 Members & Student Numbers:**

| | |
|---|---|
| Arjun Ushaben Patel | - 218196527 |
| Hanyong Zhao | - 218194050 |
| Joyce Otonglo | - 218420687 |
| Kourosh Motavalizadeh | - 217391475 |

# Table of Contents

# Executive Summary

### Problem Statement

The process of purchasing a vehicle is often a long one. For the customer shopping for a car, they would probably start by doing their research online before visiting a shop to see the vehicles and experience its features, perhaps even go on a road test. They may also hop from shop to shop in pursuit of a bargain. On the other hand, for the car dealers and salesmen, they require to engage their customers at a personal level, and take the time to explain the options to customers. A lot of time, effort and energy is spent by the car sales people in attending to customers, yet this does not in any way translate to a guaranteed sale.

It would therefore be helpful for car dealerships to make use of a statistical model that could help them predict, to a certain degree of accuracy, the propensity of new potential customers to buy a vehicle. Such a model could help them to develop strategies for targeting customers and marketing to potential purchasers as well as giving them the required attention through the purchase process, and therefore possibly increase their sales levels and ultimately meet their sales quotas.

### Intended Analysis

We are the members of the analytics department in the head office of a large car dealership chain. The marketing department needs to understand which potential clients in their database are likely to buy a car and which ones are just there for the window shopping experience.

The department has provided data including customer ID and 3 predictors: sex, age, income, and the target class variable - purchase, taking the value 1 if the client already bought a car once in the past. To predict this class for new potential customers, we need to build a logistic regression model to predict which customers are likely to buy a car, and which ones are window-shopping and not likely to buy one.

### Methodology

Logistic regression is a statistical technique used to model the probability of discrete outcomes. When properly applied, logistic regression can yield powerful insights into what attributes are more or less likely to predict event outcome in a population of interest. The models show the extent to which changes in the values of the attributes can increase or decrease the predicted probability of event outcome. Regression methods form an important part of data analysis, and describes the relationship between the independent variable (age, sex, income) and one or more outcome variables (purchase or no purchase). We performed a rigorous evaluation on the quality of prediction by testing several models of logistic regression.

We first build a complete model with all predictors and then build models with a reduced number of predictors. For each model, we will evaluate how well it is fitted using AIC and BIC scores. Furthermore, we generated a confusion matrix and classification report for each of these models on both training and testing sets.

## Analysis

### Part 1. Data preparation and visualization

We began by reading the dataset with the list of customers and the predictor values. The data was then cleaned by deleting all the rows containing missing data. Imputation was not performed on the data set. We encoded the sex predictor, which was represented by two values - male and female - by converting these string values into categorical variables in order to perform logistic regression. This is done by passing dummy variables (1 for male and 0 for females). The columns and their pairs are then visualized to give some insights into the predictors and relationships in the data. A correlation matrix is produced to make an initial guess on the usefulness of the predictors.

After preparing and cleaning the data, we visualized the columns and their pairs. The following are the key findings:

In terms of age, the brackets with the highest number of customers are 35-40 years and 40-45 years respectively (fig. 1). The most frequently occurring income bracket is $70,000- $80,000, followed by $50,000- $60,000 (fig. 2). There are slightly more female customers than males in total (fig.3).

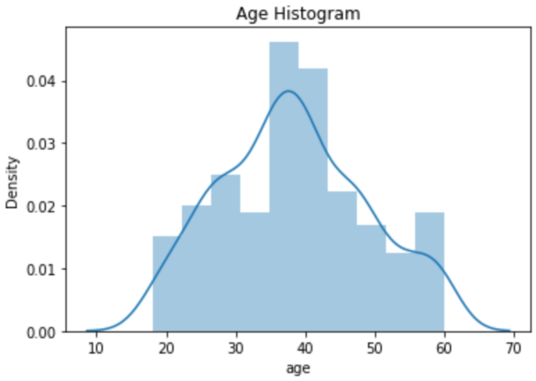


*Figure 1: Age histogram*

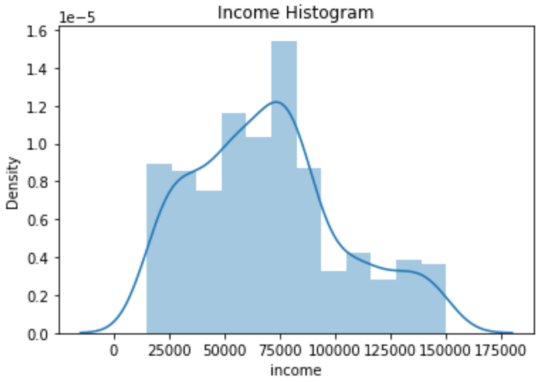


*Figure 2: Income histogram*

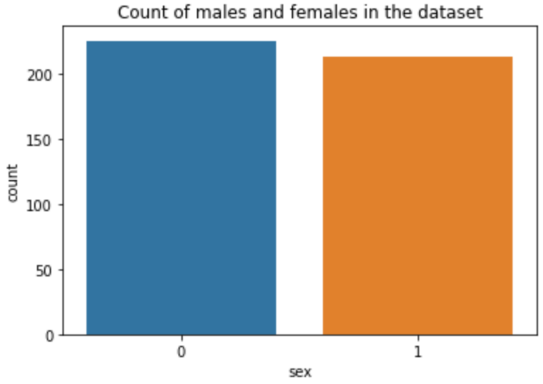


*Figure 3: Count of gender - males & females (0=female, 1=male)*

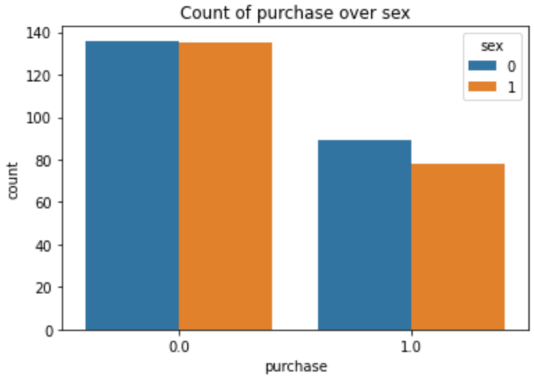


*Figure 4: Purchase by gender (0=female, 1=male)*

Amongst customers who purchased vehicles in the past, we established the following: there are slightly more women than men (fig.4); and the age with the highest number of purchases is about 46 years (fig.5).

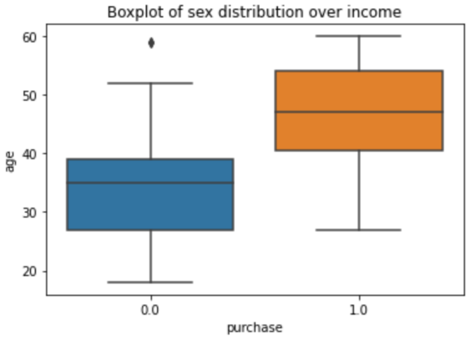
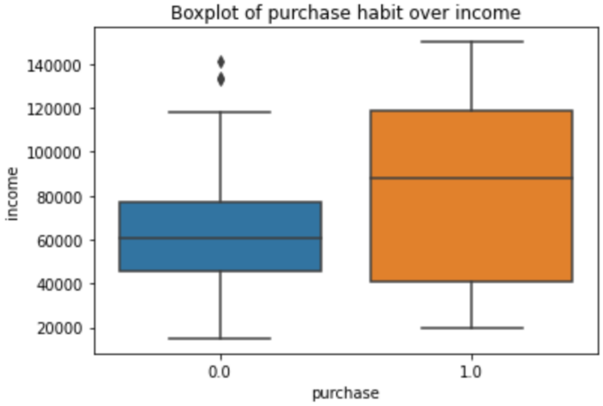


*Figure 5: Purchase by age*

Comparing purchase status and income, those who purchased vehicles have a wider income range and a higher mean salary ($90,000) than those who did not purchase vehicles ($60,000) - (fig.6). A comparison of sex and age reveals that female customers have a bigger range of age (about 31-47) than male customers (about 30-42), as well as a slightly higher mean age(fig.7).



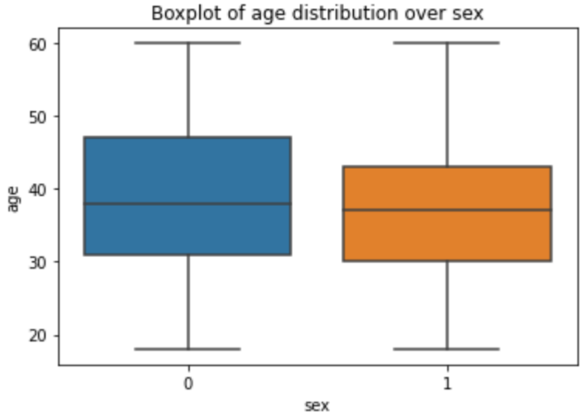


*Figure 6: Purchase by income*

*Figure 7: Sex by age (0=female, 1=male)*

A correlation matrix (see figure-1) was generated to show the usefulness of the predictors. It shows a relatively high correlation between age and purchase predictors, followed by income and purchase predictors. It is observed that age has a correlation value of 0.63 and income has a correlation of 0.34. Interestingly, the predictor sex has a negative correlation with purchase. By a naive assumption, based on this correlation matrix, we can say that age and income contribute the most for purchase of a car at the dealership.

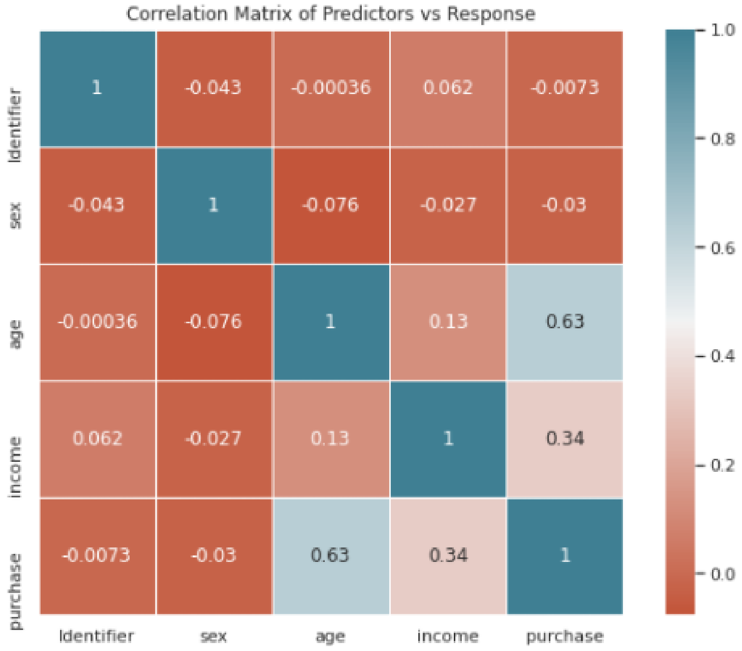


*Figure 8: Correlation between predictors and response (purchase)*

### Part 2. Inference by logistic regression

Before fitting models, we rescale the income variable so that it will produce useful coefficient results.

1. Fitting the logistic regression model with all 3 predictors results in the following regression equation:

$$Purchase = -12.7992 + 0.2946 \times sex + 0.2334 \times age + 0.0384 \times income$$

There is a 34% more chance of buying a car for male compared to females, while holding all other variables constant. An increase of 1 year in age is associated with an increase of 26% in the odds of buying a car, while holding all other variables constant. An increase of $1000 in income is associated with an increase of 4% in the odds of buying a car, while holding all other variables constant.

The AIC and BIC scores for this model are 225, 240. Something worth mentioning is that the p-values for both age and income is less than 0.001. We think both income and age are statistically significant and indicate strong evidence that there is a relationship between these two and purchase.

For sex, its p-value is 0.391, thus we think sex is not statistically significant and indicates strong evidence that there is no relationship between sex and purchase.

2. Fitting the logistic regression model with age and income results in the following regression equation:

$$Purchase = -12.4762 + 0.2301 \times age + 0.038 \times income$$

An increase of 1 year in age is associated with an increase of 26% in the odds of buying a car, while holding all other variables constant. An increase of $1000 in income is associated with an increase of 4% in the odds of buying a car, while holding all other variables constant.

The AIC and BIC scores for this model are 223, 235. Again, p-values for both age and income are less than 0.001. We think both income and age are statistically significant and indicate strong evidence that there is a relationship between these two and purchase.

3. Fitting the logistic regression model with sex and income results in the following regression equation:

   $$Purchase = -2.2518 + (-0.1427) \times age + 0.0251 \times income$$

   There is a 13% less chance of buying a car for male compared to females, while holding all other variables constant. An increase of $1000 in income is associated with an increase of 2.5% in the odds of buying a car, while holding all other variables constant.

   The AIC and BIC scores for this model are 366, 378. P-value for income is less than 0.001. We think income is statistically significant and indicates strong evidence that there is a relationship between income and purchase. For sex, its p-value is 0.577, thus we think sex is not statistically significant and indicates strong evidence that there is no relationship between sex and purchase.

4. Fitting the logistic regression model with sex and age results in the following regression equation:

   $$Purchase = -7.9219 + 0.0687 \times age + 0.1849 \times income$$

   There is a 7% more chance of buying a car for male compared to females, while holding all other variables constant. An increase of 1 year in age is associated with an increase of 20% in the odds of buying a car, while holding all other variables constant.

   The AIC and BIC scores for this model are 271, 283. P-value for age is less than 0.001. We think age is statistically significant and indicates strong evidence that there is a relationship between age and purchase. For sex, its p-value is 0.823, thus we think sex is not statistically significant and indicates strong evidence that there is no relationship between sex and purchase.

5. Fitting the logistic regression model with only age results in the following regression equation:

   $Purchase = -7.8707 + 0.1844 \times age$

   An increase of 1 year in age is associated with an increase of 20% in the odds of buying a car.

   The AIC and BIC scores for this model are 269, 277. P-value for age is less than 0.001. We think age is statistically significant and indicates strong evidence that there is a relationship between age and purchase.

6. Fitting the logistic regression model with only sex results in the following regression equation:

   $Purchase = -0.4106 + (-0.21) \times sex$

   There is a 19% less chance of buying a car for male compared to females.

   The AIC and BIC scores for this model are 408, 415. For sex, the p-value is 0.376, thus we think sex is not statistically significant and indicates strong evidence that there is no relationship between sex and purchase.

7. Fitting the logistic regression model with only income results in the following regression equation:

   $Purchase = -2.3250 + 0.0251 \times income$

   An increase of $1000 in income is associated with an increase of 2.5% in the odds of buying a car, while holding all other variables constant.

The AIC and BIC scores for this model are 365, 372. P-value for income is less than 0.001. We think income is statistically significant and indicates strong evidence that there is a relationship between income and purchase.

In summary, comparing AICs and BICs for all models, the model with predictors age and income has the lowest AIC and BIC, which means that this model is considered the best among all these models.

### Part 3. Prediction and Evaluation on different models

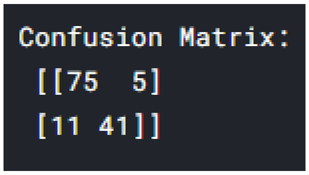
In our prediction, different models were included with varying numbers of predictors. The accuracy and confusion matrix were two metrics used in order to better understand how good the classification is performing. Also, the same methodology was used for the test set and the results are analyzed.The following table summarizes these results.

| Model | Accuracy | |
|---|---|---|
| | Train Set | Test Set |
| 1. All 3 predictors | 83% | 89% |
| 2. Dropping Sex | 83% | 88% |
| 3. Dropping Age | 78% | 75% |
| 4. Dropping Income | 82% | 87% |
| 5. Includes only Age | 82% | 87% |
| 6. Includes only Sex | 62% | 61% |
| 7. Includes only Income | 78% | 74% |

*Table 1 : Accuracy of different models on train and test set rounded at 2 decimal places*

The model has prediction accuracy range from 62% to 82% in the training set. However, analyzing the AIC, BIC and p-values of each category it showed that the customer's sex doesn't play a statistically significant role in prediction of the model. Furthermore, statistically more observations have better accuracy and this is also shown when predicting with 3 parameters. The results shows that the prediction with 3 parameters is as accurate as the one with dropping the customer's sex however, the second approach will usually be chosen for computational complexity reasons. Furthermore, the model was tested on the test set and the results were interesting in this part. The test set shows a better accuracy of prediction for the 3 parameters model which proves the above statement. Statistically speaking this makes sense however based on our analysis dropping the customer's sex variable (which is statistically negligible) should have similar or close prediction. The 3 parameters model shows a promising result of 89% and also, the model which we drop the customer's sex have accuracy of around 88%. This shows that however the 3 parameters model is showing a promising result in prediction of the test set, the model which dropped the customer's sex will be used as it is closely accurate and computationally more beneficial.

The confusion matrix of model-2 (dropping sex) will give a deeper insight on it's performance:



```
Confusion Matrix:
 [[75  5]
 [11 41]]
```

*Figure 9: Confusion matrix of the model which dropped customer's sex*

- There are 41 transactions in which the model predicted the customer will make a purchase and the customer actually purchased. These are the True Positives.
- There are 75 transactions in which the customer did not make a purchase and the model predicted the same. There are True Negatives.

- There are 11 False Negatives, i.e. the model predicted that customers will make a purchase and in reality they did not.
- Lastly, there are 5 transactions in which the model predicted a purchase but the customer did not purchase in reality. These are the False Positives.

# Conclusion

To summarize, based on our analysis, we recommend using the logistic regression model which dropped the customer's sex. The reason behind this recommendation is that statistically customer's sex variable has no significant effect in the prediction. This evidence can be further strengthened by the correlation matrix in figure-1. Before building the models, the correlation matrix suggested that age and income are the only variables which are highly correlated to purchase and customer's sex was negatively correlated.

The marketing team at the dealership needs to make sure that they don't miss out on any potential buyer and would expect the model to detect as many potential buyers as possible. For this reason, the recommended model has a higher recall than precision.

# References

1. **Dataset on Kaggle:** https://www.kaggle.com/statistics101guy/assignment-business-apps

2. **Notebook on Kaggle:** https://www.kaggle.com/arjunpatel95/mmai-5090-ai-business-apps-summer-2021-group-7