Executive Report

Assignment 2: Model Building and Analysis

ALY6040 – Data Mining Applications

Northeastern University

Professor Justin Grosz

04/23/21

Introduction:

IceCubed held a fundraising to raise money for an instant ice cream maker product. At least $100 was requested from each donor. A dataset is created by recording the information about each donor and who purchased the device. As a Data Analyst, we must conduct exploratory data analysis and develop a model to learn more about the donors who purchased the gadget, as well as the statistical importance of donor data we have with us.

Data Preprocessing and Cleaning.

The dataset consists of 10000 rows and 12 variables. Out of 12 variables 5 are numerical and 7 are Categorical data. On checking for NA values and Outliers we found issues with the Amount column where re dropped 40 values with Amount 1 (since the minimum donation amount is $100) from the dataset hence we don't perform Data cleaning for missing values and but do perform for outliers.

	Donate ID	Deposit Amount	Ice Cream Products Consumed Per Week	How many desserts do you eat a week
count	9960.000000	9960.000000	9960.000000	9960.000000
mean	4995.284739	140.076004	4.963153	6.686747
std	2886.451756	80.156594	3.165697	2.463075
min	1.000000	100.000000	0.000000	0.000000
25%	2490.750000	100.000000	2.000000	5.000000
50%	5002.500000	100.000000	5.000000	7.000000
75%	7496.250000	120.000000	8.000000	9.000000
max	9997.000000	400.000000	10.000000	10.000000

Figure 1: Descriptive Stats of Numerical Data

From Figure 1 we can deduce that the Maximum Deposit Amount is $400 and 100 as minimum. On an Avg donors consume 5 ice cream with 0 as minimum and 10 as maximum. On an Avg donors consume 7 desserts a week with 0 as minimum and 10 as maximum.

	Donate Date	Gender	Preferred Color of Device	Favorite Flavor Of Ice Cream	Donated To Kick Starter Before	Household Income	Do you own a Keurig
count	9960	9960	9960	9960	9960	9960	9960
unique	7	2	6	5	2	4	2
top	7/7/2019	male	silver	swirl	yes	Not Reported	yes
freq	3011	5254	1695	2058	6725	4781	8103

Figure 2: Descriptive Stats of Categorical Data

From the Figure 2 we can deduce that in the Gender column we have 2 unique values with male having most frequency. The Color of Device column has 6 unique values with silver color having highest frequency. In the Favorite Flavor column swirl is observed to have highest frequency.

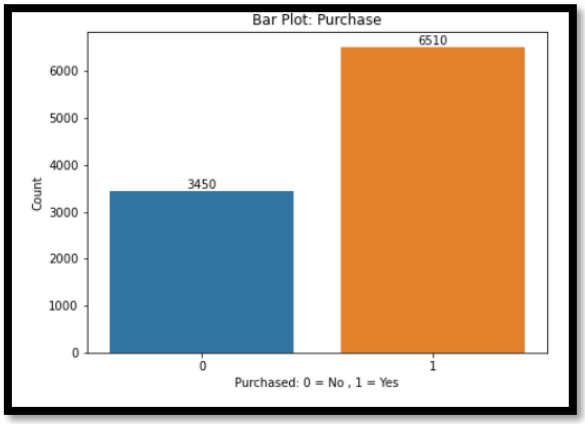

Figure 3

The Figure 3 shows count of Target variable "Purchased" for our analysis. This column tells us who purchased the device, where 6.5K donors bought the device and 3.4K didn't buy and we will check its significance with other variables.

Data Visualization:

1) The figure 4 helps us understand the purchase of machine based on donor's household income. The donors who did not report their household income have highest count of buying and not buying the device. The people with household income of <100K and >100K have purchased 1754 and 1602. People with income <50K do not look to be interested in buying the Ice cream maker.

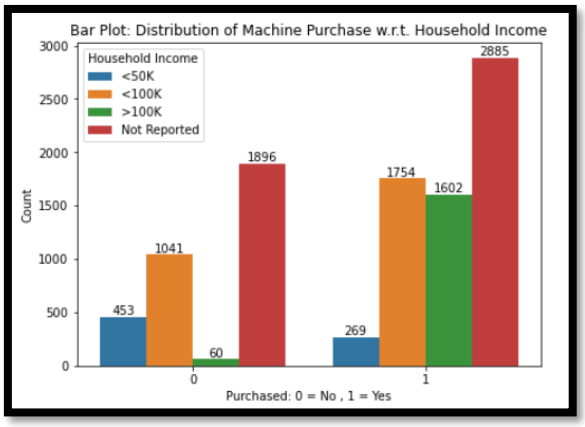

Figure 4

2) The figure 5 below helps us understand that people who have donated for the Kick Starter are interested in buying the ice cream maker. Out of almost 6K previous donor's 4.5K donors are interested in buying the device. 2.1K donors are new who donated to most recent fundraising and are interested in buying the device.

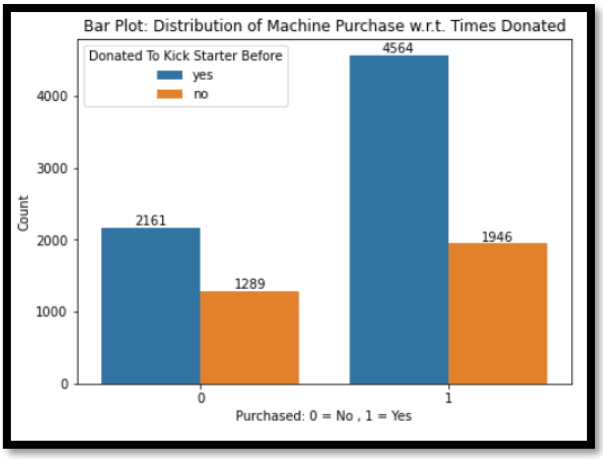

Figure 5

3) The Figure 5 depicts the distribution of people who purchased the Device and their preference of the colors of device. It is clear from the image below that people who

have registered their colors of preference as no preference have also not bought the Device.

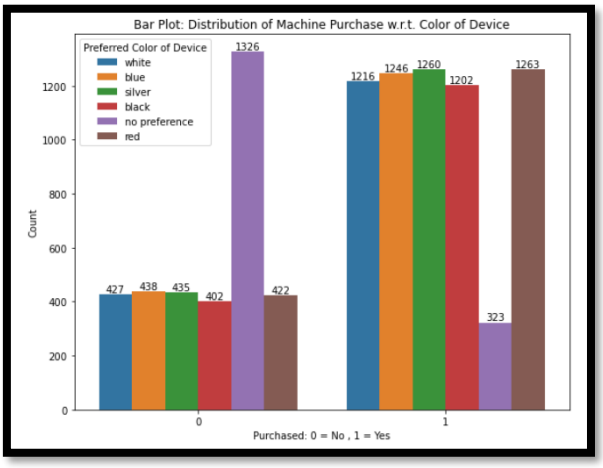

Figure 6

4) The heat map for correlation study. For the correlation study we had 4 categorical column which we encoded using the 'get_dummies' from pandas' library so that they can be included in the study. We also used 'drop_first' setting it to True to avoid the dummy trap.
Based on below correlation study in figure 7 we understand that the Purchased column has positive correlation with household income of the donor >100K which is

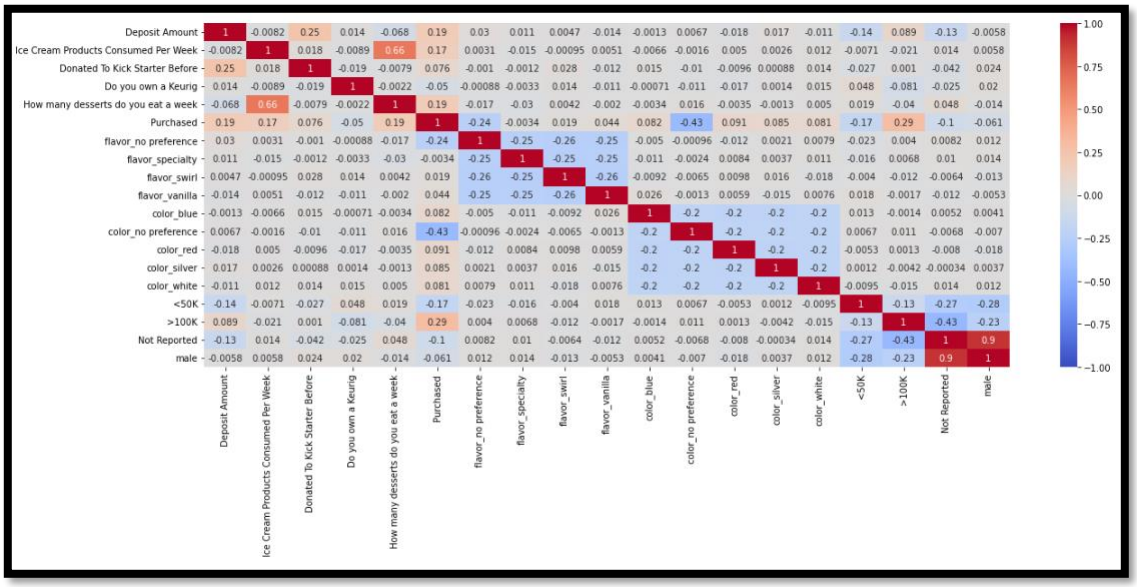

Figure 7

0.29 times. Which means when the household income is >100K we can expect the donor to purchase the device.
It can also be observed that people with no colours' preference for the device negative correlation of -0.43 with the purchased variable which we observed in the above figure 6.

Model Analysis:

When you wish to model the event probability for a categorical response variable with two outcomes, logistics regression is the best option. (*What Is Logistic Regression? | IBM*, n.d.) Hence, we will implement the Logistics Regression to predict and see what and how other variables affect the target variable.
We divide the dataset in 80:20 ratio with 80% data as the train dataset and 20 as test.

```
Optimization terminated successfully.
         Current function value: 0.330698
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:              Purchased   No. Observations:                 9960
Model:                          Logit   Df Residuals:                     9942
Method:                          MLE   Df Model:                           17
Date:               Sun, 24 Apr 2022   Pseudo R-squ.:                  0.4874
Time:                       19:32:52   Log-Likelihood:                -3293.7
converged:                      True   LL-Null:                       -6426.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
==========================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Deposit Amount                0.0121      0.001     21.816      0.000       0.011       0.013
Ice Cream Products Consumed Per Week  0.1025      0.013      7.814      0.000       0.077       0.128
Donated To Kick Starter Before  0.2855      0.065      4.380      0.000       0.158       0.413
Do you own a Keurig          -0.2470      0.078     -3.175      0.001      -0.399      -0.095
How many desserts do you eat a week  0.2355      0.016     15.079      0.000       0.205       0.266
flavor_no preference         -4.0762      0.124    -32.996      0.000      -4.318      -3.834
flavor_specialty             -2.3185      0.115    -20.217      0.000      -2.543      -2.094
flavor_swirl                 -2.1852      0.114    -19.162      0.000      -2.409      -1.962
flavor_vanilla               -1.8449      0.114    -16.210      0.000      -2.068      -1.622
color_blue                   -0.2426      0.099     -2.447      0.014      -0.437      -0.048
color_no preference          -5.5913      0.167    -33.538      0.000      -5.918      -5.265
color_red                    -0.1705      0.099     -1.723      0.085      -0.364       0.023
color_silver                 -0.1881      0.099     -1.896      0.058      -0.383       0.006
color_white                  -0.1504      0.100     -1.506      0.132      -0.346       0.045
<50K                         -1.3918      0.115    -12.112      0.000      -1.617      -1.167
>100K                         7.2645      0.292     24.870      0.000       6.692       7.837
Not Reported                  3.8546      0.330     11.666      0.000       3.207       4.502
male                         -3.8721      0.324    -11.945      0.000      -4.507      -3.237
```

Figure 8: Model Summary

Variable Significance: The p-value marked in red helps us understand the predictor significance w.r.t. the target variable. All the variables are significant enough in context to target variable apart from the device color preference of 'color_red', color_silver and 'color_white' marked in blue rectangle.

Coefficient Analysis: The coefficient of the model predictors helps us understand the change in our target variable either positive or negative based on the values obtained in our model.

a) For every change in Deposit Amount, the odds that donor will Purchase the machine are minutely on the 'Yes' side.
b) For every change in Ice Cream products consumed per week, the odds that the donor will purchase the device are 0.1 times on the 'Yes' side.
c) For the selection of Yes or No in the Donated to Kick Starter, the odds that the donor will purchase the device are 0.25 times on the 'Yes' side.
d) For every selection of Yes or No in the Do you own a Keurig, the odds that the donor

are less likely to purchase the device.

e) For every change in how many desserts do you eat a week, the odds that the donor will purchase the device are likely towards Yes.
f) Donors with flavour and device colour preferences are less likely to Purchase the device based on their coefficients.
g) Donor with household income <50K, tend to not buy the device and people with household income >100K and not reported are very likely to purchase the device.
h) Male donors tend to not purchase the device.

Confusion Matrix:
From the Confusion Matrix below we can conclude that the model we have designed is 86% accurate in predicting who will purchase the device.

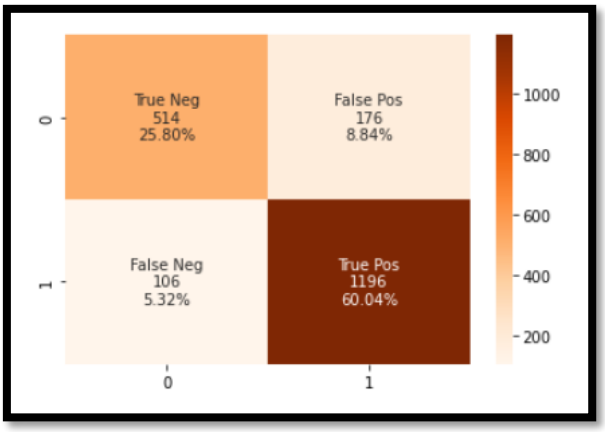

Figure 9: Confusion Matrix

- There are 514 out of 2000 people predicted that won't buy the device.
- There are 176 out of 2000 people who were expected to not buy the device but bought it.
- There are 106 people who were expected to buy the device but didn't buy it. We need to focus on this.
- There are 1196 out of 2000 people who were expected to buy the device and bought it.

Conclusion and Recommendations:

The goal of our analysis was to understand the significance of various predictor variables with the purchase of the device. During our exploration of data in correspondence to the predictors and target variables we found following insights. 65% of the donors purchased the device. To achieve our goal, we divided the dataset into training and test data set which will also help us evaluate the model. The dataset was spilt in 80:20 with training dataset as 80% and reaming as 20%. People with household income <100K and >100K have bought the device. This can also be proved from the correlation plot and from our Logistics Regression Model. People who had previously donated tend to buy the device. People with coloured device preferences have bought the device but as per our model it has negative influence during prediction analysis, and it shows they are less likely to buy the device. From the

confusion matrix we see that there are 106 donors who were expected to buy the device but didn't buy it.

Recommendation:
To increase the sale of Devices we must ask the Ice Cubed firm to target the people with household income <100K and >100K along with people who have already donated to the company during fundraiser as they have shown a good interest in buying the device during our EDA and Model Analysis.

From my analysis and understanding there are 5% donors (106 Donors from predicted confusion matrix) who were expected buy the ice cream makers didn't buy it. This could be because they already own a Keurig or were lost interested in buying the ice cream maker. It would be interesting to have a survey with these people and understand why they didn't buy the device.

From the study we understand that the male donors are less likely to buy the device, hence here too a survey within the donors group who didn't buy the device would help the Ice Cubed company to understand why even after donating to the cause the people didn't buy the device.

References:

- Bhor, Y. (2021, October 14). *Guide for building an End-to-End Logistic Regression Model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/
- T, D. (2021, December 11). *Confusion Matrix Visualization - Dennis T*. Medium. https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea
- Dobilas, S. (2022, February 5). *Logistic Regression in Python— A Helpful Guide to How It Works*. Medium. https://towardsdatascience.com/logistic-regression-in-python-a-helpful-guide-to-how-it-works-6de1ef0a2d2
- *What is Logistic regression? | IBM*. (n.d.). Logistic Regression. Retrieved April 24, 2022, from https://www.ibm.com/topics/logistic-regression
- Benton, J. (2021, December 15). *Interpreting Coefficients in Linear and Logistic Regression*. Medium. https://towardsdatascience.com/interpreting-coefficients-in-linear-and-logistic-regression-6ddf1295f6f1