Project 2: IceCubed Donor Information Analysis Report

Pei-Yu Jheng

ALY6040 – Data Mining Applications, Northeastern University

Professor Justin Grosz

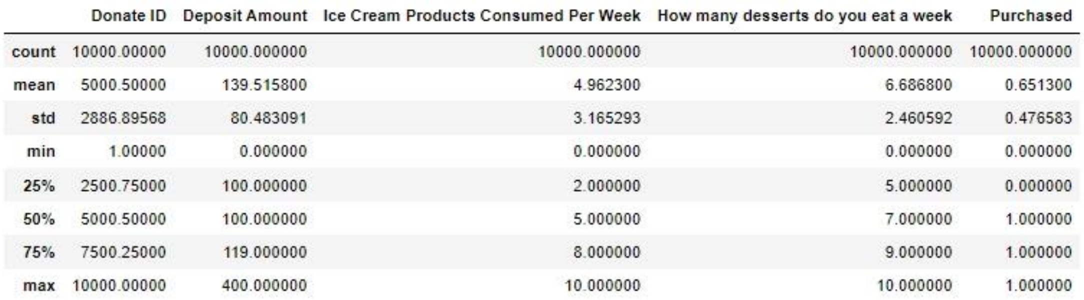
March 10, 2024

Introduction

The launch of IceCubed, a new household ice cream maker by a company, marks an exciting venture into the home appliance market. This report presents an analysis of donor data to understand the factors influencing purchase decisions for IceCubed. By examining the characteristics and behaviors of donors who purchased IceCubed compared to those who did not, we aim to provide insights that can inform marketing strategies and improve customer engagement.

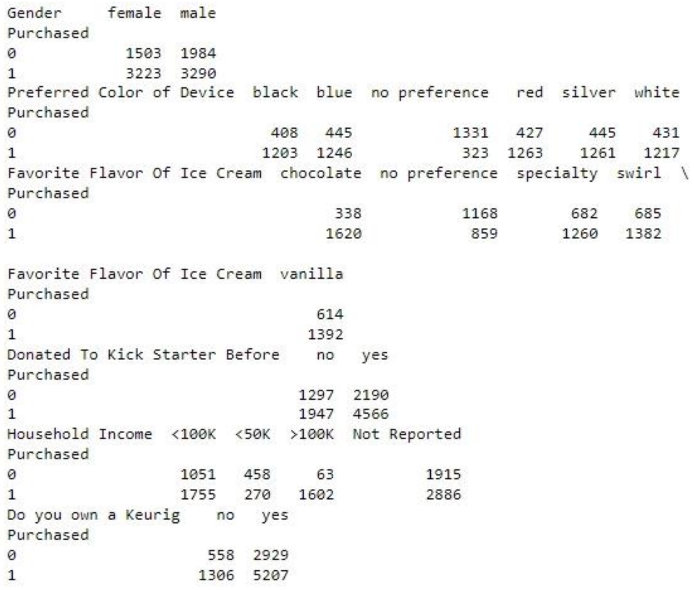
Data Cleaning

Table 1



	Donate ID	Deposit Amount	Ice Cream Products Consumed Per Week	How many desserts do you eat a week	Purchased
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.50000	139.515800	4.962300	6.686800	0.651300
std	2886.89568	80.483091	3.165293	2.460592	0.476583
min	1.00000	0.000000	0.000000	0.000000	0.000000
25%	2500.75000	100.000000	2.000000	5.000000	0.000000
50%	5000.50000	100.000000	5.000000	7.000000	1.000000
75%	7500.25000	119.000000	8.000000	9.000000	1.000000
max	10000.00000	400.000000	10.000000	10.000000	1.000000

Figure 1



```
Gender     female  male
Purchased
0            1503  1984
1            3223  3290
Preferred Color of Device  black  blue  no preference   red  silver  white
Purchased
0                            408   445           1331   427     445    431
1                           1203  1246            323  1263    1261   1217
Favorite Flavor Of Ice Cream  chocolate  no preference  specialty  swirl  \
Purchased
0                                   338           1168        682    685
1                                  1620            859       1260   1382

Favorite Flavor Of Ice Cream  vanilla
Purchased
0                                 614
1                                1392
Donated To Kick Starter Before    no   yes
Purchased
0                               1297  2190
1                               1947  4566
Household Income  <100K  <50K  >100K  Not Reported
Purchased
0                  1051   458     63          1915
1                  1755   270   1602          2886
Do you own a Keurig    no   yes
Purchased
0                     558  2929
1                    1306  5207
```

The interpretation of the cleaning process results is as follows:

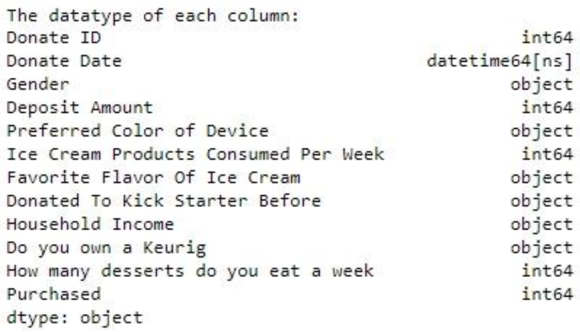
1. Summary Statistics:

- According to Table 1 and Figure 1, we can see that there are no outliers in the dataset.

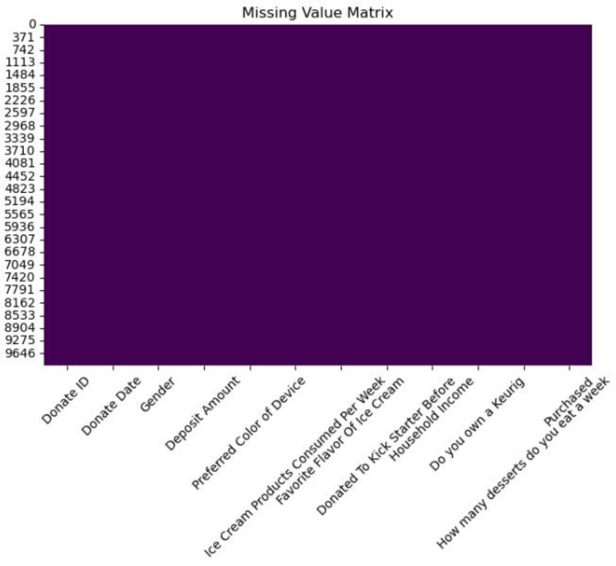
- Deposit Amount: The mean deposit amount is approximately $139.52, with a standard deviation of $80.48. The minimum deposit amount is $0, and the maximum is $400.

- Ice Cream Products Consumed Per Week: The mean number of ice cream products consumed per week is approximately 4.96, with a standard deviation of 3.17. The range is from 0 to 10.
- How many desserts do you eat a week: The mean number of desserts eaten per week is approximately 6.69, with a standard deviation of 2.46. The range is from 0 to 10.
- Purchased: The mean indicates that around 65.13% of respondents purchased the product.

Figure 2



```
The datatype of each column:
Donate ID                                         int64
Donate Date                              datetime64[ns]
Gender                                           object
Deposit Amount                                    int64
Preferred Color of Device                        object
Ice Cream Products Consumed Per Week              int64
Favorite Flavor Of Ice Cream                     object
Donated To Kick Starter Before                   object
Household Income                                 object
Do you own a Keurig                              object
How many desserts do you eat a week               int64
Purchased                                         int64
dtype: object
```



2. Data Types:

- “Donate ID”, ”Deposit Amount”, ”Ice Cream Products Consumed Per Week”, ”How many desserts do you eat a week”, and ”Purchased” are all numerical columns.
- ”Donate Date”, ”Gender”, ”Preferred Color of Device”, ”Favorite Flavor Of Ice Cream”, ”Donated To Kick Starter Before”, ”Household Income”, and ”Do you own a Keurig” are categorical columns. We encoded them as numbers.

3. Missing Values:

- There are no missing values in any of the columns, as the heatmap visualization confirms that there are no missing values, as there are no visible gaps in the heatmap.

4. Duplicate and irrelevant values

- There are 0 duplicates in the dataset.
- We dropped the irrelevant feature "Donate ID" and "Donate Date" since the date starts from July 1st to 9th 2019.

5. Feature engineer 2 new variables:

- Desserts Consumed: This calculation is done to get the number of desserts consumed excluding ice cream. The assumption here is that the total number of desserts consumed may provide additional information that could be useful for predicting the target variable.
- Interaction Desserts and Income: This variable could capture the interaction effect between the frequency of consuming ice cream products and the household income level. It could be calculated by multiplying the "Ice Cream Products Consumed Per Week" by a numeric representation of the "Household Income" (We encoded the income categories as numbers).

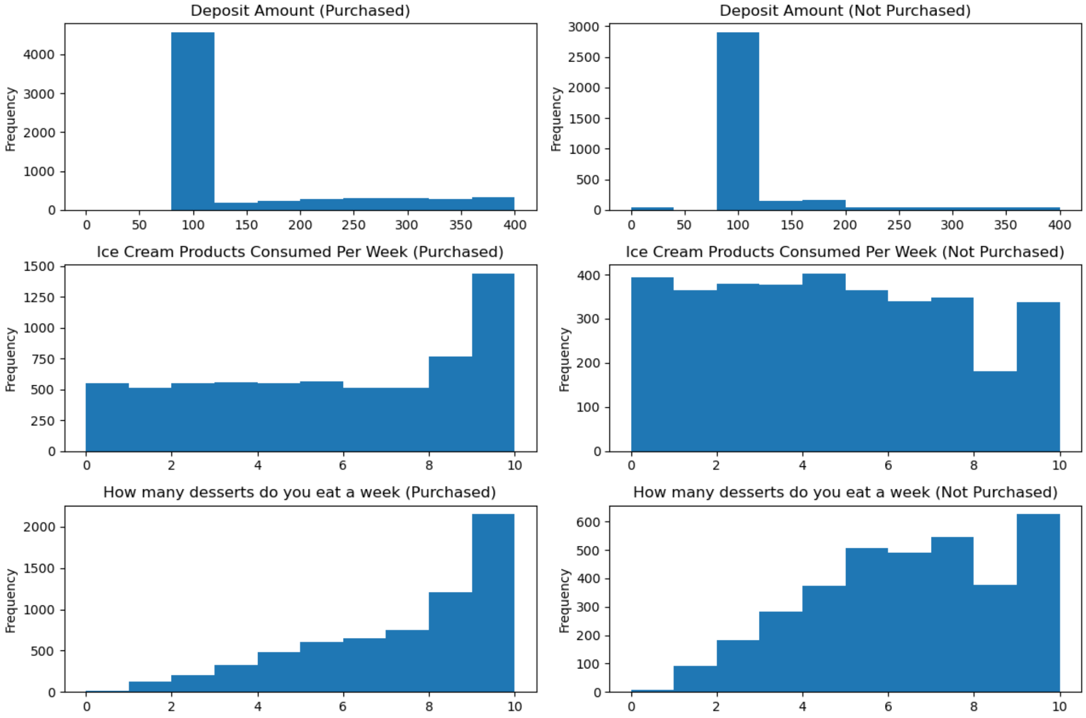
These results provide a comprehensive overview of the dataset's characteristics, including summary statistics, outliers, data types, missing values, check duplicates, irrelevant features, and categorical variable encoding, which are essential for further data analysis and modeling.

Problem Background

Understanding customer behavior is essential for companies to make informed decisions. By analyzing the data, we can gain insights into customer preferences and purchasing behavior. This information can help IceCubed improve its marketing strategies and target potential customers more effectively. This analysis utilizes donor data collected during the pre-launch phase of IceCubed. The data includes demographic information, purchase history, and survey responses. We will use statistical analysis techniques, including exploratory data analysis (EDA) and hypothesis testing, to identify significant differences between the two groups of donors.

Q1: Distribution of numerical variables

Figure 3



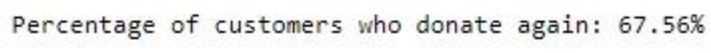
Insights from EDA:

- $100 was the popular choice for donors. Understanding the popularity of certain donation amounts can also help the company forecast its revenue more accurately. It provides insight into expected donation levels and can inform budgeting and planning decisions. In addition, personalize fundraising appeals based on past donation amounts. For donors who have previously donated $100, tailor the message to thank them for their past support and encourage them to continue their generosity.

- Customers who purchased the product tend to consume more ice cream products and eat more desserts per week.

Q2: How many percent of customers donate again?

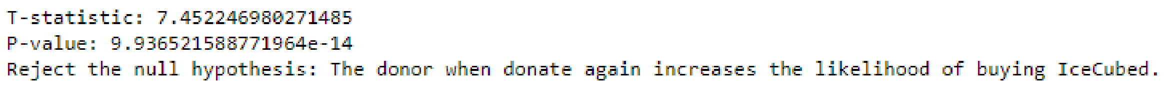
Figure 4



```
Percentage of customers who donate again: 67.56%
```

***Will the donor who donated increase the likelihood of buying IceCubed? (T-test)*¶**

Figure 5

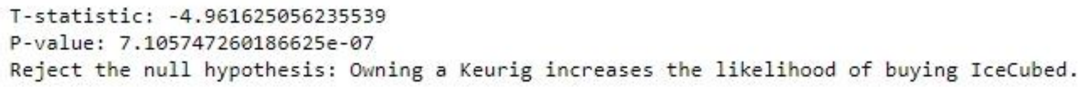


```
T-statistic: 7.452246980271485
P-value: 9.936521588771964e-14
Reject the null hypothesis: The donor when donate again increases the likelihood of buying IceCubed.
```

For the marketing team, the percentage of customers who donate again (67.56%) indicates a high level of customer loyalty and repeat engagement. This metric suggests that a significant portion of customers who have donated once are likely to donate again in the future. The T-test also proved it. Understanding this behavior can guide the marketing team in developing strategies to nurture and maintain customer relationships.

Q3: If owning a Keurig increase the likelihood of buying IceCubed? (T-test)

Figure 6

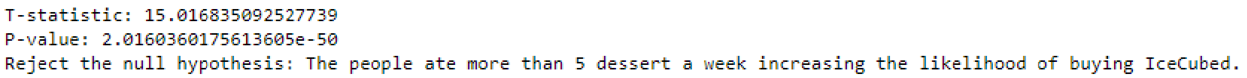


```
T-statistic: -4.961625056235539
P-value: 7.105747260186625e-07
Reject the null hypothesis: Owning a Keurig increases the likelihood of buying IceCubed.
```

The company can use this information to target Keurig owners with marketing campaigns or promotions, as they are more likely to purchase IceCubed. It may also indicate potential market segments or customer profiles that are more receptive to IceCubed, allowing for more targeted marketing strategies. Understanding this relationship can help the company allocate resources more effectively and focus its efforts on the most promising customer segments.

Q4: Did the consumption of dessert increase the likelihood of buying IceCubed? (T-test)

Figure 7



```
T-statistic: 15.016835092527739
P-value: 2.0160360175613605e-50
Reject the null hypothesis: The people ate more than 5 dessert a week increasing the likelihood of buying IceCubed.
```

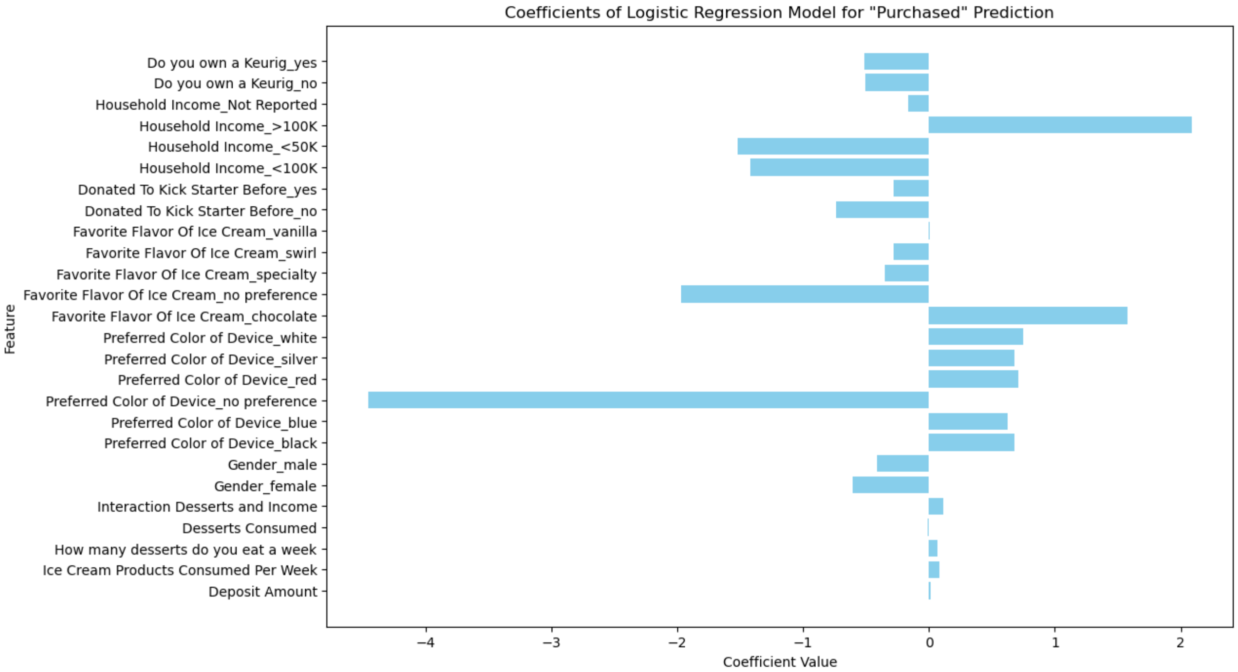
We reject the null hypothesis, which means that consuming more than 5 desserts a week does increase the likelihood of buying IceCubed. This information could be valuable for the marketing team, as it suggests a potential target demographic for the product.

Model Development and Evaluation

In order to reveal the relationships between the purchased variable and other features, we conducted a logistic regression model. When preparing the train and test set, we split the 0 and 1 in the "Purchased" variable equally to prevent imbalanced leads to the result misleading. Below are the results of the model.

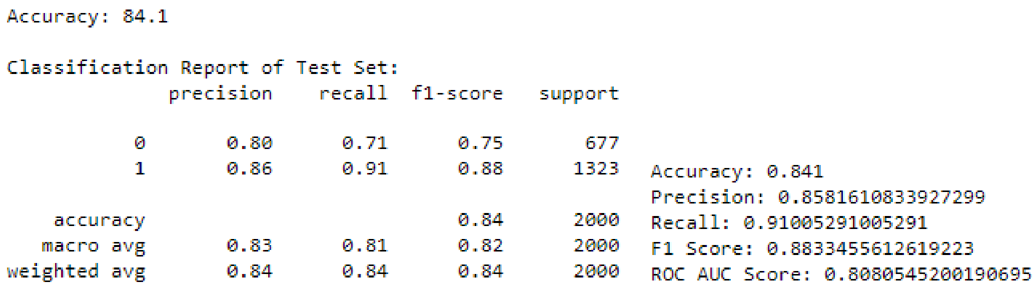
Figure 8

Visualization of the model's coefficient



Through this chart, we can clearly see these variables’ positive or negative impact on the Purchased variable. These coefficients represent the impact of each variable on the likelihood of a customer purchasing the IceCubed product. “Preferred Color of Device_no preference”, “Household Income_>100K”, and “Favorite Flavor of Ice Cream_no preference” have the most significant impact on “Purchased.” Next, we are going to evaluate our model.

Figure 9

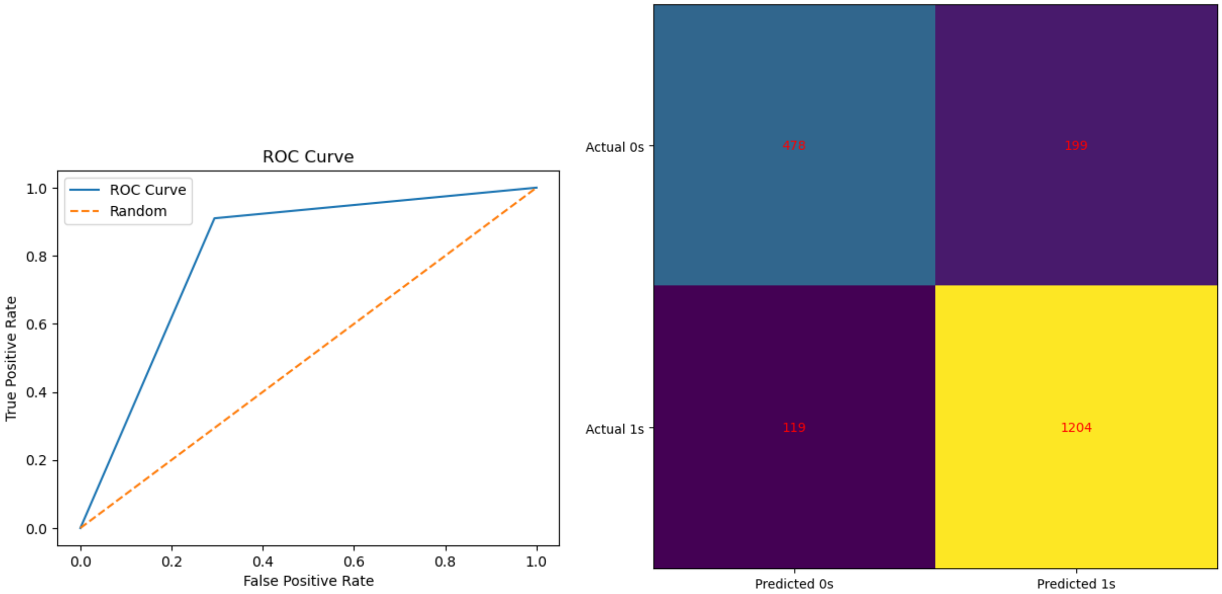


```
Accuracy: 84.1

Classification Report of Test Set:
              precision    recall  f1-score   support

           0       0.80      0.71      0.75       677
           1       0.86      0.91      0.88      1323    Accuracy: 0.841
                                                         Precision: 0.8581610833927299
    accuracy                           0.84      2000    Recall: 0.91005291005291
   macro avg       0.83      0.81      0.82      2000    F1 Score: 0.8833455612619223
weighted avg       0.84      0.84      0.84      2000    ROC AUC Score: 0.8080545200190695
```

Figure 10 – ROC Curve & Confusion Matrix



These metrics in Figures 9 and 10 provide an evaluation of how well the logistic regression model performs in predicting whether a customer will purchase IceCubed or not. An accuracy of 0.84 means that the model correctly predicts the purchase or non-purchase of

IceCubed 84% of the time. A recall of 0.91 means that the model correctly identifies 91% of all actual purchases. An AUC score of 0.808 indicates that the model has good discriminatory power in distinguishing between positive and negative instances.

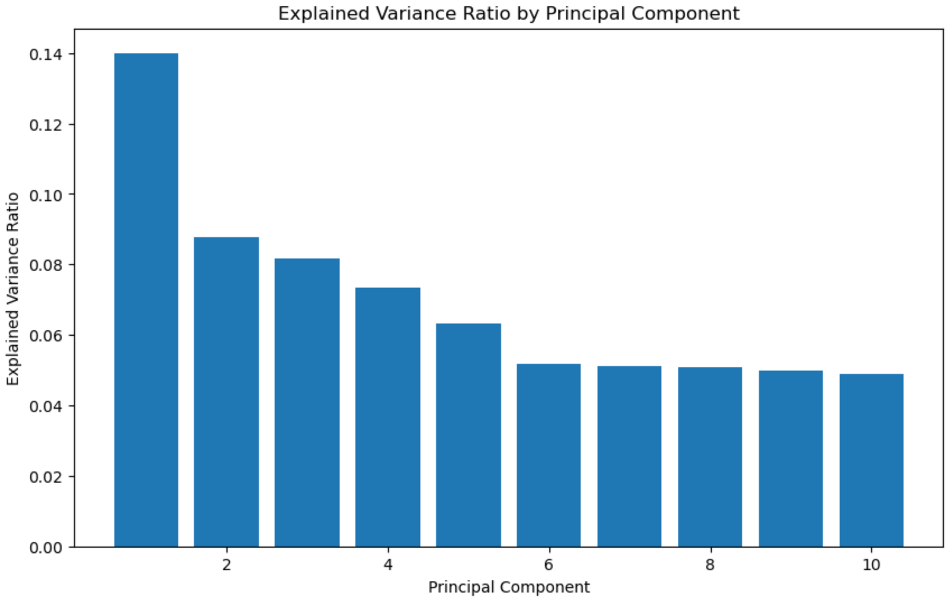
Confusion Matrix:

- True Positive (TP): 1204, instances the model correctly predicted 1204 purchases.
- False Positive (FP): 199, instances the model incorrectly predicted 199 purchases.
- True Negative (TN): 478, instances the model correctly predicted 478 non-purchases.
- False Negative (FN): 119, instances the model incorrectly predicted 119 non-purchases.

Overall, these metrics suggest that the logistic regression model performs reasonably well in predicting purchases of IceCubed, with high precision, recall, and F1 score, as well as a relatively high ROC AUC score.

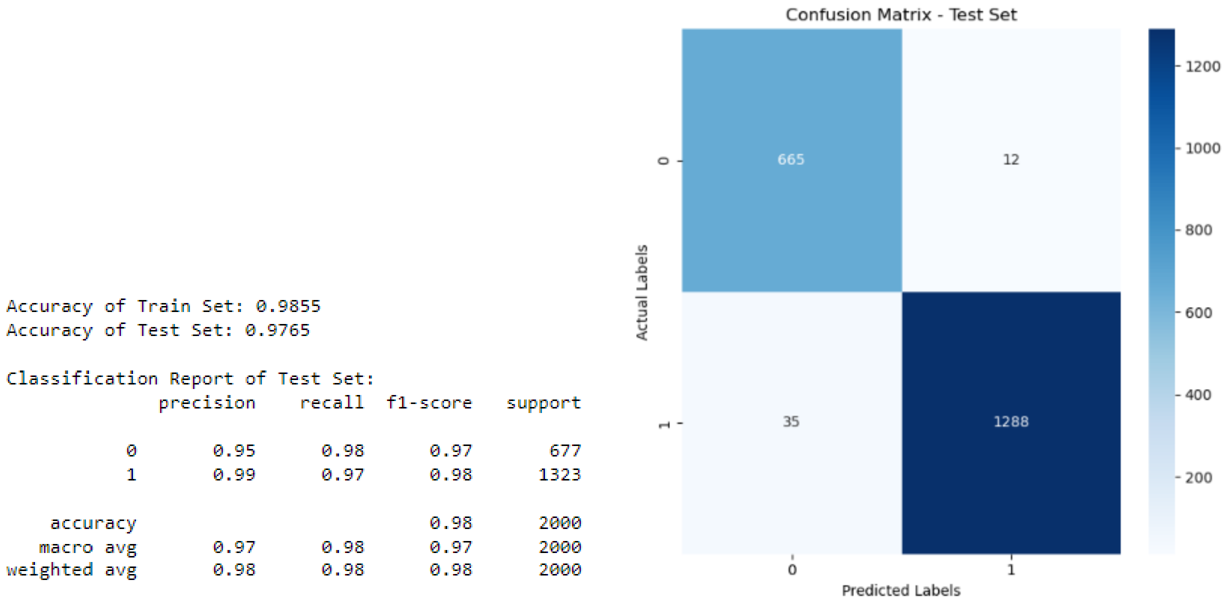
Logistic Model with PCA

Figure 11



In order to decide which PCs will help us enhance the previous model, we plot a char of Explained Variance Ratio by Principal Component, which identifies PC5 as an inflection point where the rate of decline in explained variance ratios becomes more consistent. By choosing PC5, you aim to balance capturing a substantial amount of variance while minimizing information loss. PC5 represents a point where the additional variance explained by additional components is relatively small.

Figure 12



```
Accuracy of Train Set: 0.9855
Accuracy of Test Set: 0.9765

Classification Report of Test Set:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97       677
           1       0.99      0.97      0.98      1323

    accuracy                           0.98      2000
   macro avg       0.97      0.98      0.97      2000
weighted avg       0.98      0.98      0.98      2000
```

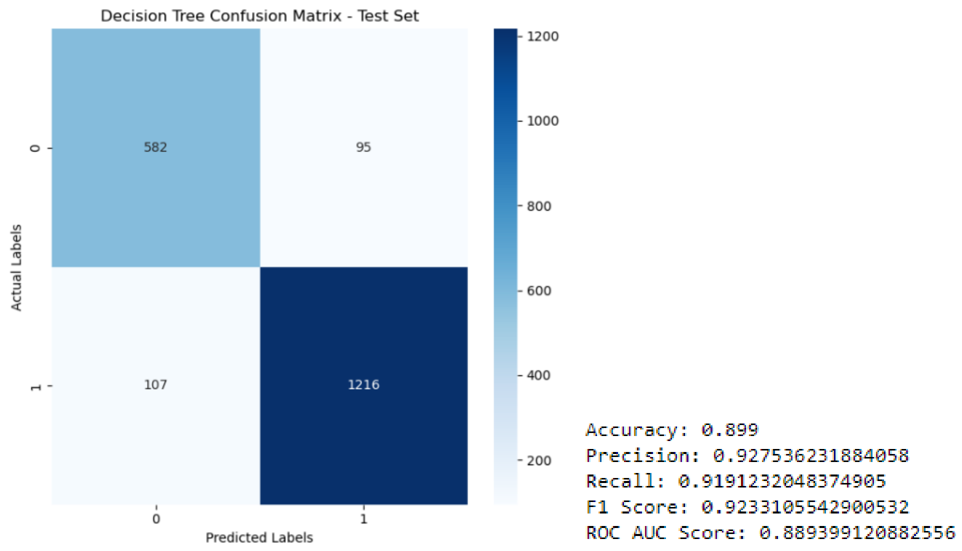
Using PCA as a preprocessing step before applying logistic regression has significantly improved the model performance. Here's a summary of the improvements:

- The accuracy has increased from 0.841 to 0.9765, indicating that the model's overall correctness has greatly improved.
- The ROC AUC score, which measures the area under the receiver operating characteristic curve, has increased from 0.808 to 0.978, indicating a significant improvement in the model's ability to distinguish between positive and negative classes.
- The confusion matrix shows a substantial reduction in false positives (from 199 to 12) and false negatives (from 119 to 35), indicating that the model's predictions are more accurate after using PCA.

Overall, using PCA as a preprocessing step has led to a more accurate and reliable logistic regression model for the given dataset.

Decision Tree

Figure 13

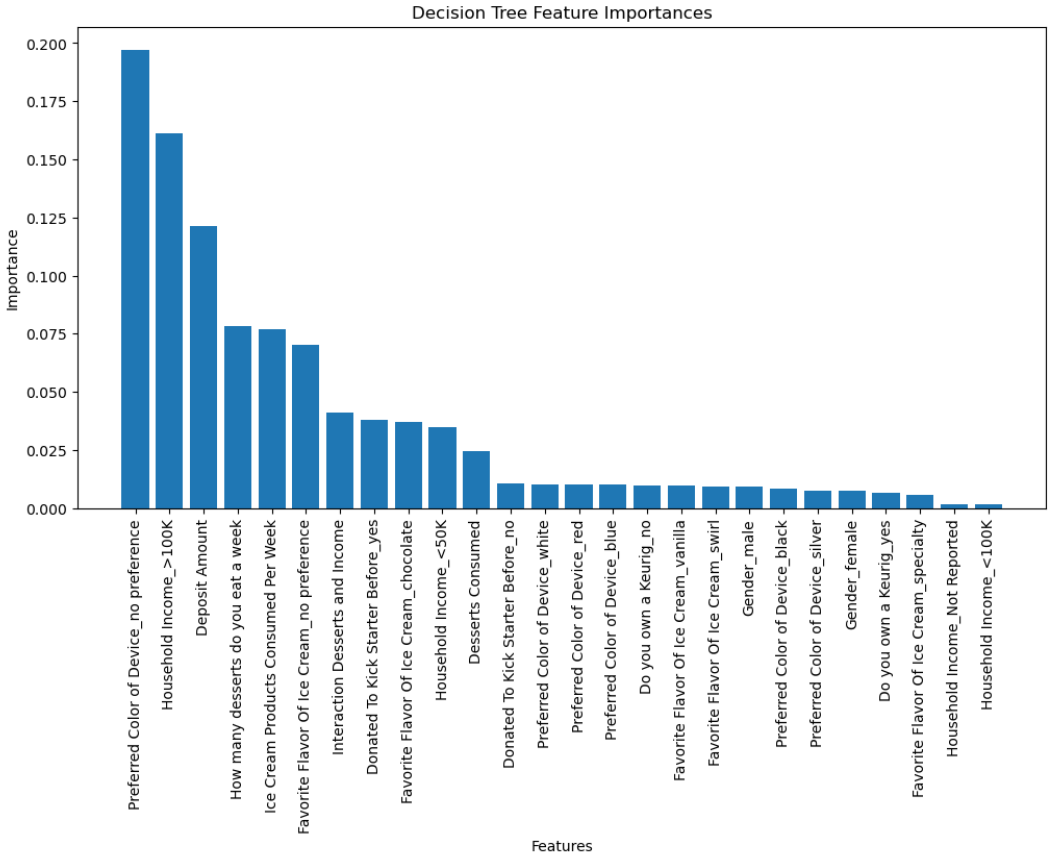


```
Accuracy: 0.899
Precision: 0.927536231884058
Recall: 0.9191232048374905
F1 Score: 0.9233105542900532
ROC AUC Score: 0.889399120882556
```

The metric in Figure 13 provides an evaluation of how well the decision tree performs in predicting whether a customer will purchase IceCubed or not. An accuracy of 0.899 means that the model correctly predicts the purchase or non-purchase of IceCubed almost 90% of the time. A recall of 0.919 means that the model correctly identifies 91.9% of all actual purchases. An AUC score of 0.889 indicates that the model has good discriminatory power in distinguishing between positive and negative instances.

Figure 14 - Feature Importance

For the decision tree, the feature importances show the importance of each feature in the decision tree model. Features with higher importance are more influential in predicting the target variable. We can see that the top 3 features that influence the target variable are “Preferred Color of Device_no preference”, “Household Income_>100K”, and “Deposit Amount.” The top 1 and the top 2 in Figure 14 are the same with the Logistic Model; however, the decision tree identified the "Deposit Amount" as top 3 not "Favorite Flavor of Ice Cream_no preference."



Compare the results of these models

Table 2

	Precision	Recall	F-1 score	Accuracy	False positive	False negative
Logistic Regression	0.85	0.91	0.88	0.84	199	119
Logistic Regression with PCA	0.99	0.97	0.98	0.97	12	35
Decision Tree	0.92	0.91	0.92	0.89	95	107

1. Logistic Regression vs. Logistic Regression with PCA: Using PCA as a preprocessing step before logistic regression improved the precision, recall, and F1 scoreThe accuracy also

increased. There was no increase in false positives or false negatives, indicating that the PCA transformation helped improve the model's performance without increasing misclassifications.

2. Logistic Regression with PCA vs. Decision Tree: The logistic regression model with PCA outperformed the decision tree in terms of precision, recall, and F1 score. It also had higher accuracy. However, the decision tree had fewer false positives and false negatives, suggesting that it may generalize better to unseen data.
3. Decision Tree vs. Logistic Regression: The decision tree outperformed the logistic regression model in terms of precision, recall, and F1 score. It also had a slightly higher accuracy. However, there was no increase in false positives or false negatives, indicating that the decision tree model did not show signs of being more prone to misclassification than the logistic regression model.

Overall, the Logistic Regression with PCA model outperforms the other models in terms of precision, recall, F1 score, and accuracy. It also has the lowest number of false positives and false negatives, making it the most reliable model among the three.

Analysis of Variable Significance

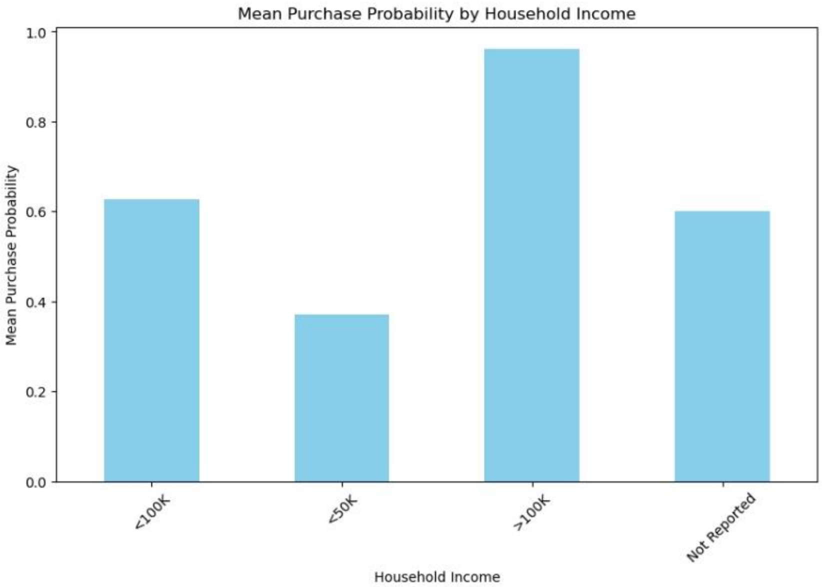
The coefficients of the Logistic Model in Figure 8, here's what they mean to the company:

1. Deposit Amount, Ice Cream Products Consumed Per Week, How Many Desserts Do You Eat a Week, Desserts Consumed, Interaction Desserts and Income: These variables have positive coefficients, indicating that as these values increase, the likelihood of a customer purchasing IceCubed also increases. The company can use this information to target customers who are likely to purchase based on their spending habits and consumption patterns. It aligns with EDA previously.

2. Gender_female, Gender_male: These variables have negative coefficients, suggesting that gender may not have a significant impact on purchasing behavior.
3. Preferred Color of Device: The coefficients for different color preferences indicate varying degrees of impact on purchasing behavior. For example, customers who prefer black, blue, red, or white devices are more likely to purchase IceCubed compared to those with no color preference. However, customers with no color preference have a significantly lower likelihood of purchasing.
4. Favorite Flavor of Ice Cream: Customers who prefer chocolate ice cream are more likely to purchase IceCubed, while those with no preference or who prefer specialty or swirl flavors are less likely to purchase it.
5. Donated to Kickstarter Before: Customers who have donated to Kickstarter before are less likely to purchase IceCubed compared to those who have not donated before.
6. Do You Own a Keurig: Customers who do not own a Keurig are less likely to purchase IceCubed compared to those who do own a Keurig.

We can see the largest positive coefficient was “Household Income >100k.” For the negative coefficients, the highest was “Preferred Color of Device – no preference,” and the second one was “Household Income <50k;” therefore, we want to know the mean purchase probability by Household Income.

Figure 15



Higher household income (>100K) is associated with a higher likelihood of purchasing IceCubed, while lower income levels (<100K, <50K) are associated with a lower likelihood. Customers who did not report their household income also have a lower likelihood of purchasing.

The top 6 feature importances of the Decision Tree in Figure 14, here's what they mean to the company:

1. Preferred Color of Device_no preference: This feature has the highest importance; however, we cannot know if it indicates that customers who do not have a color preference are more likely to purchase IceCubed or not. We only know it affects our target variable the most.
2. Household Income_>100K: Customers with a household income greater than $100,000 are the second most important feature, implying that higher-income customers are more likely to purchase IceCubed. The top 1 and top 2 are the same with Logistic Regression.
3. Deposit Amount: The deposit amount also plays a significant role, suggesting that customers who are willing to make a larger initial investment are more likely to purchase. This variable is more significant in Decision Tree Model.

4. How many desserts do you eat a week and Ice Cream Products Consumed Per Week: These features indicate that customers who consume more desserts and ice cream products per week are more likely to purchase IceCubed, which is more significant in Decision Tree.
5. Favorite Flavor Of Ice Cream_no preference: Customers with no specific favorite flavor of ice cream are less likely to purchase IceCubed, which aligns with the findings from logistic regression.
6. Interaction Desserts and Income: There is an interaction between the number of desserts consumed and income, suggesting that this interaction can impact the likelihood of purchase. This variable is more significant in the Decision Tree.

Differences:

- Logistic regression coefficients indicate the direction and strength of the relationship between each feature and the likelihood of purchase. For example, a positive coefficient for a feature means that an increase in that feature's value increases the likelihood of purchase, while a negative coefficient means the opposite.
- Decision tree feature importance provides a relative ranking of features based on their ability to split the data and predict the outcome. It does not provide information about the direction or strength of the relationship.

Business Focus and Recommendations

The analysis of coefficients reveals significant variables that influence purchasing behavior. Here are the interpretations:

1. The analysis reveals several key insights into customer behavior and preferences regarding IceCubed. The positive coefficients for variables such as Income, Deposit Amount, Ice Cream Products Consumed Per Week, and How Many Desserts Do You Eat a Week indicate

that customers who have a household income of more than a hundred thousand, spend more on deposits, consume more ice cream products, and desserts per week are more likely to purchase IceCubed. This suggests that targeting customers with higher income and higher spending habits in these areas could lead to increased sales of IceCubed.

2. The negative coefficients for variables like Gender, Preferred Color of Device (no preference), Favorite Flavor of Ice Cream (no preference, specialty, swirl), and Donated to Kickstarter Before suggest that these factors have a lesser impact on purchasing behavior. However, it's important to note that some color and flavor preferences do influence purchasing behavior positively, indicating that certain product attributes can still be leveraged in marketing strategies.

Recommendations:

- Target High-Income and High-Spending Customers: Focus marketing efforts on customers who have income of more than a hundred thousand and exhibit higher spending habits, such as those who make larger deposits and consume more ice cream products and desserts per week. Tailor messaging and promotions to highlight the benefits of IceCubed that appeal to these customers' preferences and lifestyles.
- Leverage Positive Influencers: Identify and target customers who exhibit positive purchasing behaviors, such as those who prefer certain colors or flavors associated with higher purchase likelihood. Use these preferences to personalize marketing messages and promotions, potentially increasing engagement and conversion rates.
- Consider Additional Demographic Factors: While gender may not have a significant impact, other demographic factors like household income and ownership of a Keurig do influence purchasing behavior. Incorporating additional demographic variables, such as age, location,

or education level, could provide further insights into customer segments and preferences, allowing for more targeted marketing strategies.

Overall, by focusing on customers with higher spending habits and leveraging positive influencers, IceCubed can improve its marketing strategies and increase customer engagement and sales. Incorporating additional demographic variables can further enhance these strategies by providing a more nuanced understanding of customer preferences and behaviors.

This report provides a comprehensive analysis of the IceCubed donor dataset and offers valuable insights into customer behavior. Implementing the recommendations provided can help IceCubed improve its marketing strategies and increase its sales.

Reference

Data Source:

ALY6040 Data Mining Application Course of Northeastern University, https://northeastern.instructure.com/courses/165136/assignments/2071169

Python, R. (2023, June 26). *Logistic regression in Python*. https://realpython.com/logistic-regression-python/