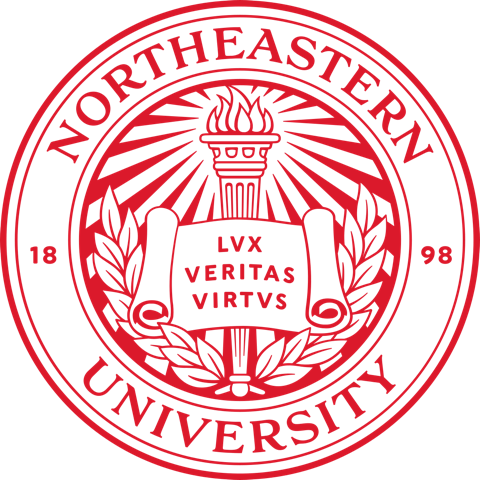
**Course:** ALY 6040

**Instructor’s Name:** Prof. Dr. Justin Grosz



**Submitted By:**

Sunil Raj Thota

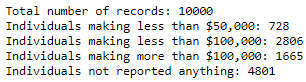
**Date:** 10/10/2021

**Introduction**

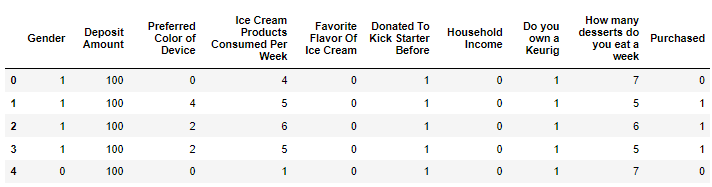
This research examines the impact of several variables on the purchase factor of a product called IceCubed, as well as the data modeling process. In this, I have looked at the associations in the information to see if there were any markers of success of the project that I could use. I have gone through various characteristics of the information and displayed the findings after a primary data cleansing step. The data cleaning and data profiling techniques from Assignment 1 were applied to Kickstarter’s data set used in this study. It is expected that the data is clean and ready for data modeling. In this, let’s dive deep into the donor's purchasing prediction using the data and understand the insights that we bring in.

**Analysis**

Now, let’s do some data cleansing and exploratory data analysis. In this process, I have found some interesting insights which might be useful in our inferences.



From the above picture, we can depict that 48% of the Individuals did not report their Household incomes.



Categorical data consists of variables with label values instead of numeric values. The number of available values is frequently restricted to a small number. This implies that category data must be transformed into numerical data. If the class label is a target value, you might like to transform the model's forecasts back to categorical form to present them or use them in a program. The integer encoding is insufficient for category variables with no such ordinal connection.

The binary variables are often called “dummy variables” in other fields, such as statistics. We got dummies for Preferred Color of Device and Favorite Flavor of Ice Cream results which results in extra columns in the dataset. To perform Logistic Regression analysis, we need to change the categorical variables to hardcoded integers. Here, we have performed this technique for Gender, Donated to Kick Starter Before, Household income, Do you own a Keurig column in the dataset.

In the same way that linear regression finds an equation that predicts an outcome for a binary variable, Y, from one or more response variables, X, logistic regression (LR) finds a solution that predicts an outcome for a binary variable, Y, from one or more dependent variable, X. The dependent variables, unlike linear regression, can be categorical or continuous, as the model does not require continuous data. When contrasted to other supervised classification methods like kernel SVM or ensemble algorithms, logistic regression is comparatively quick, but its accuracy suffers.

For this, let split the dataset into independent and dependent variables. We will be using the Logit function to fit the model with the necessary variables to see the p values and other statistics of each column.

I also went ahead and visualized the correlation between the variables using a corr heatmap. Used liblinear as the solver in the Logistic Regression method observed the model fit. And, obtaining a confusion matrix as shown below.



And then I have calculated the Accuracy, Precision, Recall, and F1-Scores metrics as shown in the below image. The ratio of correctly classified subjects to the total number of subjects is known as accuracy. The ratio of accurately +ve labeled to all +ve labeled is known as precision. A recall is the proportion of those who are in reality to those who are appropriately +ve classified. Precision and recall are both taken into account while calculating the F1 Score. It's the precision and recall's average.



Now let’s perform Decision Tree Algorithm on our dataset. It is not necessary to normalize the data while using this technique. This algorithm can also be implemented without scaling the data. It is not necessary to impute missing values in this algorithm. For decision trees, the data pre-processing step requires less code and analysis. I have dropped our target viable i.e., the Purchased column in the train and test splits. There is no optimal split percentage and I choose test size as 30% and train size as 70%. In the Decision Tree Classifier, I chose Gini as the criterion. The Gini Index, also known as Gini impurity, assesses the likelihood of a certain feature being erroneously categorized when chosen at random. The characteristics with the lowest Gini Index value would be favored while creating the decision tree.

After training the Decision tree classifier, we need to predict the model with the test data and obtained an accuracy as 0.886



Now, let’s perform Random Forest Classifier Algorithm. A random forest is a supervised machine learning system that uses decision tree algorithms to build it. This algorithm is used to anticipate behavior and results in a variety of sectors.

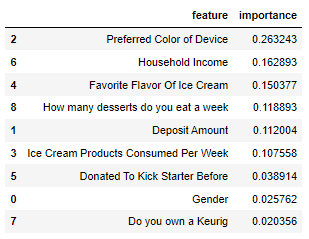
A random forest is a machine learning technique for solving classification and regression problems. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. A random forest method overcomes the drawbacks of a decision tree algorithm. It minimizes dataset overfitting and improves accuracy.

It is based on several decision trees. There are three types of nodes in a decision tree: decision nodes, leaf nodes, and root nodes. Each tree's leaf node represents the final output generated by that particular decision tree. The final product is chosen using a majority-voting procedure.

Let’s fit the model with training data and predict the probabilities and accuracy.



Let’s get the importance of certain features in the model



From the above figure, Preferred Color of Device, Household Income, Favorite flavor of Ice Cream, How many Desserts do you eat a week, Deposit Amount is the key driving factors for getting the better accuracy.

Let’s compare the results of all the models

Table 1: Models Comparison Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **Logistic Regression** | *0.827* | *0.837* | *0.911* | *0.872* |
| **Decision Tree** | *0.886* | *0.93* | *0.89* | *0.91* |
| **Random Forest** | *0.885* | *0.91* | *0.91* | *0.91* |

**Interpretation and Recommendations**

From the above analysis and interpretations, I would like to recommend IceCubed product to have more data by performing extensive surveying about their products to all the customers of all ages and regions. As this is a household ice cream maker product, they should be focusing and targeting the family audience with better household income.

I would also love to know more about the children in a family and considering it as a factor in purchasing this product. As most of the children love ice creams, it would be a great parameter to look for. To convert the donors to customers perfectly, the company has to invest some money and resources in promoting their brand by endorsing and supporting popular brands, companies, organizations, and influencers. A company should act and plan its marketing strategies to better profit from the products. Also, it would be great to know the donor’s region or city, or state and see the purchasing turnout patterns accordingly.

The preferred color of the device parameter is also important and played as a major decision-making attribute in converting a donor to purchase this product. Because device color plays an important factor in deciding a user’s purchasing behavior and interests. Also, a person’s favorite flavor of ice cream is important because while selling the IceCubed product, it is necessary to have multiple flavored options to make ice-creams. They have to do a lot of market research and competitive analysis for the same.

**Appendix: Data Visualizations**

Figure 1: Violin Plot of all Continuous variables

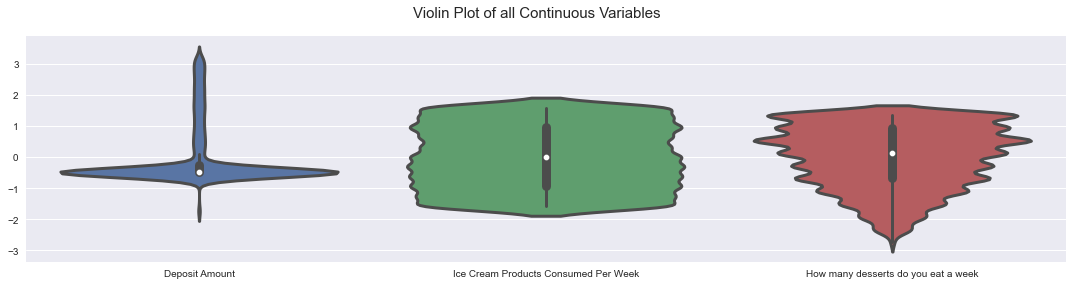
****

Figure 2: Box plots of Do you own a Keurig and donated to kick starter before or not

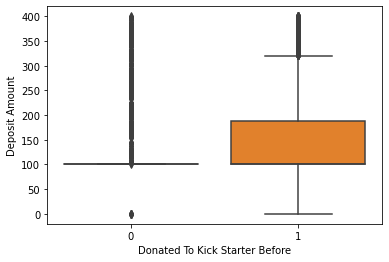
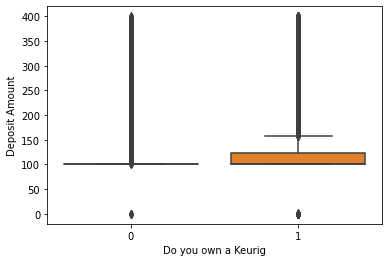
****

Figure 3: Correlation Plot

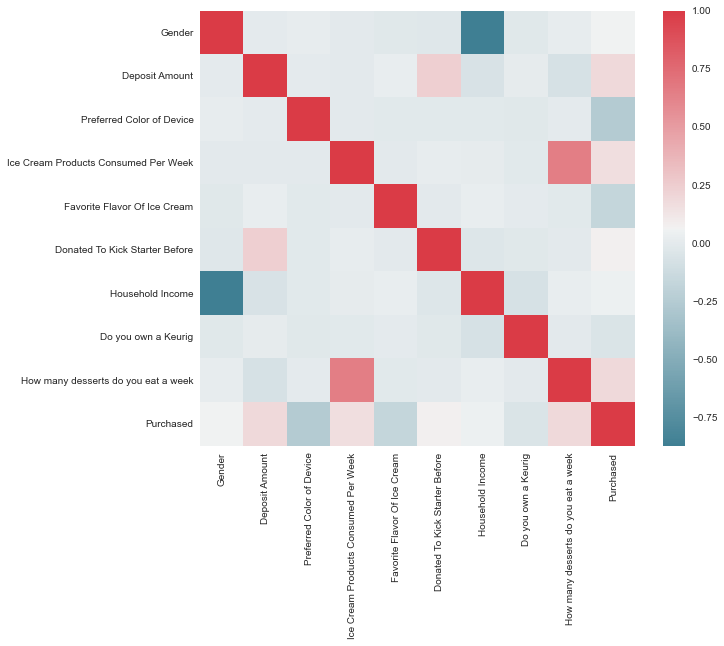


Figure 4: Correlation Plot

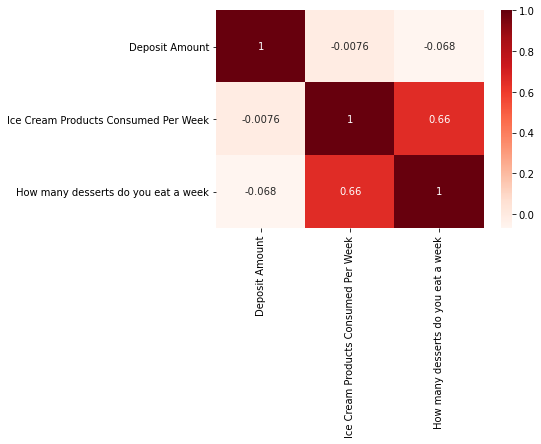


Figure 5: Visualizing Important Features

