# Technique Practice

# Predictive Modeling

# Module-2

# Akshay Kirit Kalavadia

College of Professional Studies, Northeastern University ALY 6040 – Data Mining Application

Prof. Justin Grosz

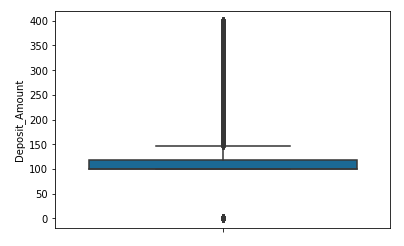
Fall 2021

# Introduction

In this project, we will look deeper into the dataset of Kickstarter’s IceCubed product. IceCubed is ice cream making machine and Kickstarter started a fundraiser for this project. The dataset contains various details regarding the product launch regarding the donor. In this project, we will analyze the data to find out important variables that drive the decision making of purchasing the product. Before diving into analysis part, we need to perform Exploratory Data analysis and data cleaning which will prepare the data for accurate analysis and results.

The dataset contains 10,000 rows and 12 columns. While checking the cleanliness of the data, it is observed that there all no Null values in the dataset. The data which doesn’t have a defined value are either “No Preference” or “Not Reported”. The numeric variables are explored using describe() function to check mean, median, mode and quartile values and to check for outliers in the dataset. To check for the outliers in categorical variables, I have used Boxplot which revealed that Donation amount has abnormal data.

**Graph – 1: Boxplot of Deposit Amount**

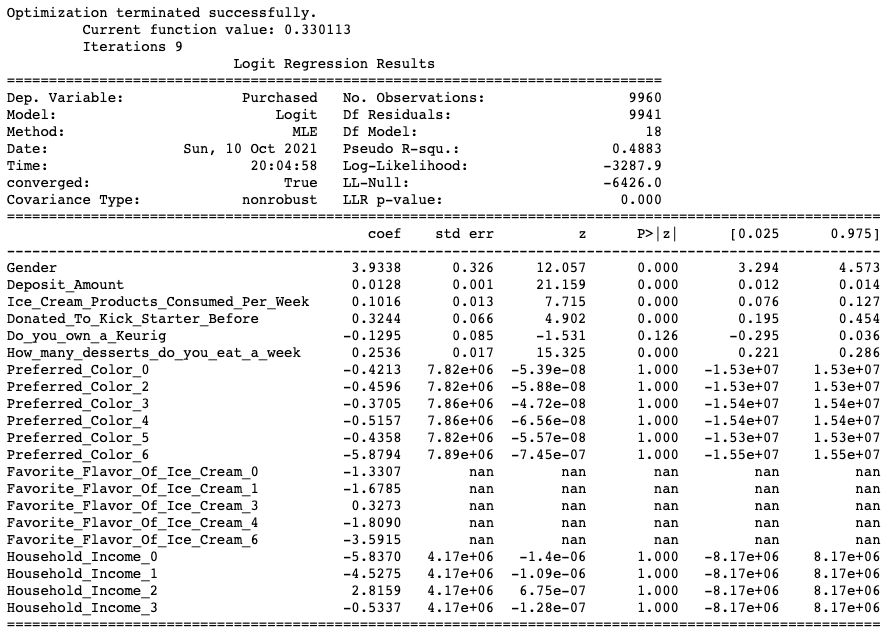


It is observed that 39 values have $0 as Donation amount while 1 value has $1 as donation amount. Its logical that a person who has not given any donation cannot be considered as Donor. I have cleaned the data for these 40 rows by removing them.

**Analysis**

The purpose of this project is to build a predictive model which can predict if a donor will purchase IceCubed based on various important variables. For this assignment, I have used Logistic Regression, Decision Tree and Random Forest methods to build Predictive models. I will judge the performance of the model based on 3 parameters, i.e., Accuracy, AIC, MSE and speed of each model. This will help us understand which model is best for Kickstarter to predict the decision making of purchasing the machine.

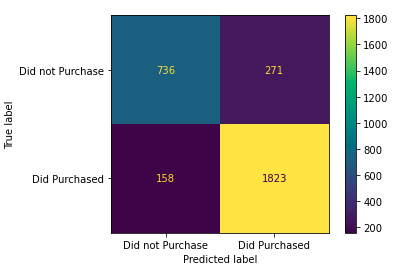
In Logistic Regression, the model is built to identify important categorical variables using One-Hot coding. From the Logistic model, I obtained following result:



From the P-value of each variable; Gender, Deposit Amount, Ice Cream Products Consumed Per Week, Donated To Kick Starter Before and How many desserts do you eat a week are important variables. Gender has the highest co-efficient value followed by Donated to Kickstarter Before and How many desserts do you eat a week. According to this model, Gender of a person is most important in predicting if he/she will purchase IceCubed machine.

The following graph shows the confusion matrix generated using Logistic Regression:

**Graph – 2: Logistic Regression Model - Confusion Matrix**



Accuracy of the model is 0.856.

AIC of the model is 6613.85.

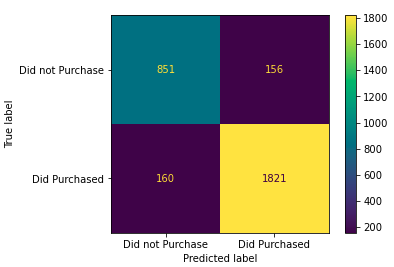
MSE of the model is 0.144.

In Decision tree model, the data is split based on certain parameters which is used to make prediction of values for the traget or dependent vaiable. We have decision nodes and leaves. Leaves are the final decision.

The Attribute Selective measure for this model is Gini index. The feature with least gini index is selected. In general terms, Gini index is a metric that measures the likelihood of a variable being incorrectly categorized when it is picked at random. As per the model, model is split based on Preferred color followed by Favourite flavor of ice cream and Household income.

Confusion matrix of this model is as follows:

**Graph – 3: Decision Tree Model - Confusion Matrix**



Accuracy of the model is 0.894.

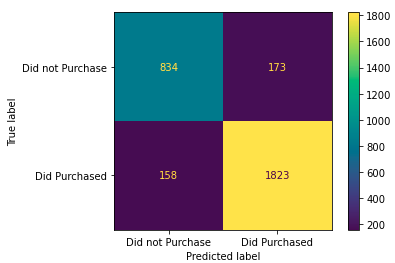
AIC of the model is 5964.49.

MSE of the model is 0.105.

Speed of the model is 0.014 sec.

In Random Forest model, the model runs multiple Decision tree models to accurately predict the dependent variable based on independent variables. Following confusion matrix shows the result after the Random Forest is performed on train and test models to predict dependent variable y.

**Graph – 4: Random Forest Model - Confusion Matrix**



From the confusion matrix, we can calculate the accuracy of the model. Following are the result for different measures from Random Forest model:

Accuracy of the model is 0.889.

AIC of the model is 5964.39.

MSE of the model is 0.110.

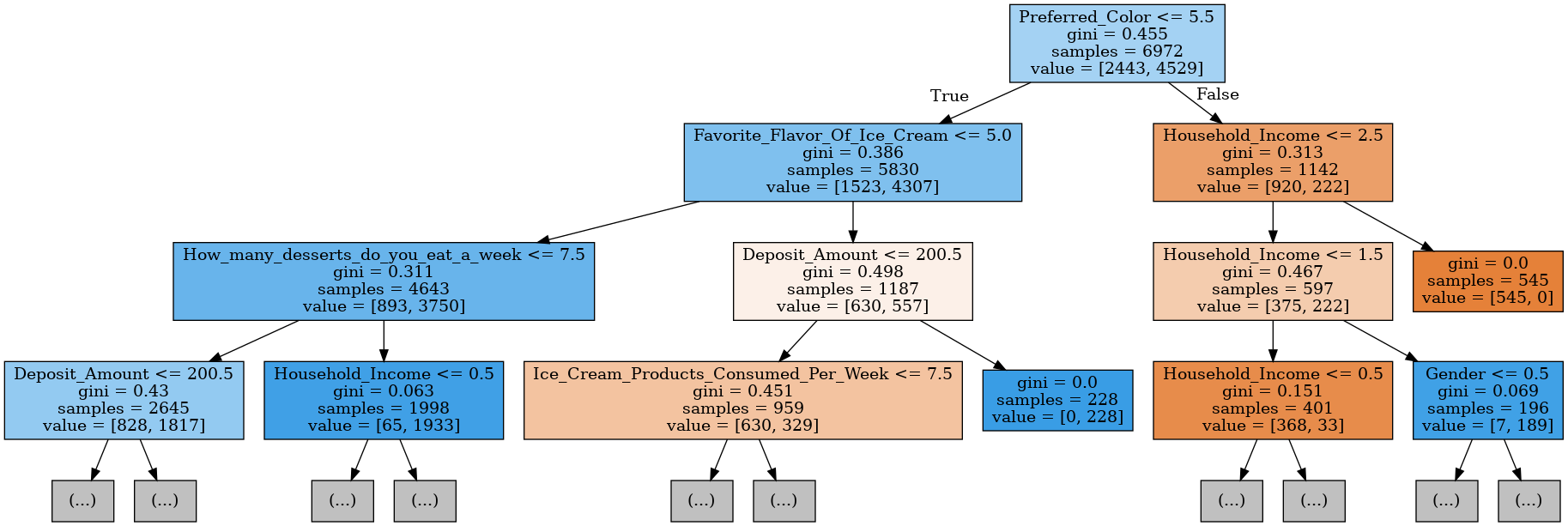
Speed of the model is 0.62 sec.

After comparing all three models, I observed that **Decision Tree Model** has the highest Accuracy, lowest AIC, lowest MSE and is fastest model of all the three models. Using this model, the manufactoring company can predict the probability of selling an IceCubed machine based on few parameters of the donors. Based on this dataset, the model can be tested with outside clients to see if they would buy the products.

**Conclusion**

Few of the most important parameters or variables for better predicition are Preferred colour of machine, household income, favourite flavour of ice cream, deposite amount, how many desserts does a person eat in a week and the number of ice cream products consumed per week. This observation is based on the feature selection technique implemented on Decision Tree Model. Following plot supports the same:

**Graph – 5: Decision Tree :**



Based on this model, the manufacturer should provide color option of the machine which are most prefered by the target customers. Once we learn about the same, most prefered color should be placed on displays and used during demo sessions. We can also infer that people with higher household income are the target clients of the company.

**References**

*1.10. decision trees. scikit*. (n.d.). Retrieved October 10, 2021, from <https://scikit-learn.org/stable/modules/tree.html>.

*Random Forest Python example from scratch ... - youtube.com*. (n.d.). Retrieved October 10, 2021, from <https://www.youtube.com/watch?v=aXw12xFVxwg>.

Navlani, A. (2018, Decemeber 28). *Python decision tree classification with Scikit-Learn Decisiontreeclassifier*. DataCamp Community. Retrieved October 10, 2021, from <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>.

Towards AI. (2020, May 28). *Decision Tree Explained With a Practical Example.* Retrieved October 10, 2021, from <https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53>