Abhigna Ramamurthy

NUID: 002982276

ALY6040: DATA MINING APPLICATIONS

Module 2: Technical Practice

March 6, 2022

Table of Contents

[Introduction 2](#_Toc97502579)

[Problem Background 2](#_Toc97502580)

[Data Clean-up 3](#_Toc97502581)

[Data Analysis and Interpretation 3](#_Toc97502582)

[Correlation Analysis 4](#_Toc97502583)

[Data Modelling 4](#_Toc97502584)

[Conclusion and Recommendations 6](#_Toc97502585)

[References 7](#_Toc97502586)

[Appendix 8](#_Toc97502587)

# Introduction

The report aims to cover analysis on Kickstarter company which has recently launched a new product called ‘IceCubed’. The data provided includes data on the campaign with target customers done along with details on whether those who participated in the campaign purchased the device or not. Customer analysis such as this have enormous applications in the real world. Predictive modelling such as one showed in this report not only helps understand the customer behavior, but also help in predicting the probability of a new customer purchasing the device. This analysis can feed into the company financial analysis to forecast growth of profits. The objective of this project is to investigate the data set to look for hidden patterns, understand the factors contributing to the decision of buying the product and build a predictive model to determine whether the customer purchased or not.

# Problem Background

The new venture of Kickstarter, called ‘IceCubed’ is like a Keurig device for ice creams. It is an innovative product that tries to change the customer behavior. Customers are generally used to buying ice cream outside or as a big tub to enjoy at home. The device helps to customize ice cream and get ice cream their favorite flavor at any time of the day. The company conducted a 9-day campaign to do initial data collection on customer behavior. The data collected includes customer’s attributes like gender, preferred color of device, favorite flavor of ice cream, whether they have donated to Kickstarter before, their household income range (<50K, <100K, >100K), and if they owned a Keurig before. Numerical attributes like their donation amount, ice cream products consumed per week and number of desserts consumed per week is also present. The team at Kickstarted have also mapped this donation data with whether the same person purchased the device or not under the ‘Purchased’ column attribute which has a Boolean value indicating 0 for not purchased and 1 for purchased. The company now wants to understand how the customer conversion rate is, i.e., how many donated and purchased. Details on customer behavior based on the preferences mentioned during the campaign can also give valuable insights to the company. They also want to build a predictive model which can efficiently predict if the customer will buy the product or not based on the details collected during the campaign.

The “kickstarter” dataset has 10,000 records and 12 attributes with details on customer preferences and if they purchased the device or not. There are 5 integer values, 6 object values which are categorical data, and 1 date time value. Each donation or record can be identified by the unique identifier of ‘Donation ID’. None of the columns have null values, or duplicate values. Insights into various categorical numerical values pertaining to the purchase of the device and modelling for a predictive analysis on the purchase is investigated in the further report.

# Data Clean-up

As there are no null values or duplicate values, analysis on validity of the data was done.

From the grouped bar chart of categorical values in Figure 1 of Appendix, we can see that dataset has people who voted for “No Preference” for preferred color of device and favorite flavor of ice cream. These are valid data as many people cannot pick only one color or flavor. Household Income column also has many records with “Not reported” value. This again can’t be considered as null values as people would have preferred not to share this data. Hence replacing it with statistical mean or median would not be a correct method to analyze the data.

Outliers in a dataset can lead to flawed analysis. Figure 2 box plot of numerical values helps identify that there are outliers present in “Deposit Amount” of the dataset. On close investigation, there are many zero Deposit Amount records which just means that the people took the survey put did not donate anything to the Kickstarter. Values higher than the 150 are also marked as outliers, but I decided to keep these values as the amounts are true values of donation and are just higher than most other donation amounts. Other numerical columns of the dataset do not have any outliers present.

# Data Analysis and Interpretation

Through the “Purchased” attribute of the dataset we can immediately tell that the 9-day campaign was a success as about 65.39% of customers who took part in the campaign ended up buying the device. This showcases that knowledge of the device and effective communication during the campaign survey resulted in the sale of the device.

First step in data analysis involves analyzing the spread of the dataset. From the grouped bar chart of Figure 3, which shows the categorical attributes distribution with respect to people who purchased and did not purchase the device. We can see that there is almost no significant distinction between number of males and female gender identified people buying the device. It is interesting to note that majority of the people who voted as no preference to “Preferred color of the device” did not buy the device. It can be said that people who took in the time to explore the different color of devices available ended up buying. We can also say that almost all colors of the device have catered to different population. However, red color has won the race of highest number of devices purchased. Another important categorical variable to analyze is “Favorite Flavor of ice cream”. People who prefer chocolate flavor have bought the highest number of devices by a huge margin. We can conclude that chocolate as the top flavor of the customers. The grouped bar chart for “Donated to Kickstarter before” shows that customer involvement in the campaign helps in sales of the product. More than half of the people who had previously donated to Kickstarter bought the device. “Household Income” gives insight on the economic background of the customers. From the grouped bar chart, we can see that highest number of devices was bought by people who did not report their income. However, from the people who reported their income, <100K group bought the highest number of devices, and the conversion rate of number people who reported >100K and ended up buying the device is very high. Hence, we can conclude that above 50K households would be the main target customers for the company. People who owned a Keurig before brought the highest number of sales for the company.

Histogram of numerical attributes of the dataset provide insight on what kind of distribution they follow. From the histogram of Figure 4, helps identify the “Deposit Amount” has a right skewed normal distribution, “Ice cream products consumed per week” has a uniform distribution and “How many desserts do you eat per week” has linearly increasing values. From the bar chart of Figure 5, we can see that more the number of ice cream products consumed per week, higher the chances of buying the device. This conclusion can also be confirmed by bar chart of Figure 6, which shows that higher the consumption of dessert products per week, higher chances of buying the device.

## Correlation Analysis

Correlation analysis is a well-known method to quantify association between two continuous values. It produces the “Pearson Product Moment correlation coefficient” which indicates the direction and strength of linear association.

Chart

Description automatically generatedFigure 7: Correlation heat map of Kickstarter dataset

From the heatmap on the left, we can see that Purchased has positive correlation with other numerical values of the dataset. Donation Id can be ignored as it won’t be considered for the final analysis.

Using One-hot encoding method, I created dummy variables of categorical variables are created to analyze the effects on Purchased. This helps convert a categorical variable to numerical variable which is easy for the algorithm to include. The correlation plot from Figure 8 includes the correlation co-efficient of categorical variables along with numerical attributes. We can see that “Household Income\_>100K”, “Deposit Amount”, “How many desserts do you eat per week”, and “Favorite Flavor Of Ice Cream\_chocolate“ are the top four positively correlated attributes to “Purchased”. The top four negatively correlated attributes are “Household Income\_Not Reported”, “Household Income\_<50K”, “Favorite Flavor Of Ice Cream\_no preference”, “Preferred Color of Device\_no preference”.

# Data Modelling

The aim of data modelling in this report is to accurately find whether the customer purchased the device or not. Correlation analysis above provided an outline of what attributes contribute to the purchase of “IceCubed”, based on that, there are three models I used to do predictive analysis on the purchase. As the target is binary, logistic regression is the obvious choice. I split the data into training (70%) and test data (30%) for modelling purpose. I used the training data to perform logistic regression analysis which resulted in 85.06% accuracy, 87.6% precision and 90.4% recall. The Figure 9 shows the confusion matrix for the first model with logistic regression which provides summary of results of the classification problem. 60.30% of true positive, i.e., the model predicted purchased and the customer purchased the device, 24.77% true negative, i.e., the model predicted not purchased, and the customer did not purchase, 8.53% false positive, i.e., the model predicted that the customer did not make the purchase, but they did, and finally, 6.4% false negative which is the model predicted that the customer would buy but they did not. Overall, the false negatives need to be reduced to positively affect the sales of the device. The figure 10 shows the ROC (Receiver Operator Characteristic) of the model as a comparison between an ideal model and the model I created.

The co-efficients identify “Household Income\_>100K” as the most influential factor and “Preferred color of device\_no preference” as the highest negatively affecting factor for the prediction of “Purchased”. The R-squared for the model is 0.486 and AIC is 4721.9601. These values will help in comparison of the model with other two models explained below.

The next model I implemented is Recursive Feature Elimination (RFE), is a feature selection algorithm. It helps in eliminating the features that are not necessary. RFE function helped to identify the following attributes to be eliminated: “Deposit Amount”, “Ice cream products consumed per week”, “Favorite flavor of ice cream\_swirl” and “Do you own a Keurig\_no”. The new dataset was again split into training (70%) and testing (30%). I used the training data to perform logistic regression on the reduced features which resulted in 84.4% accuracy, 86.1% precision and 91.35% recall. The Figure 11 shows the confusion matrix for the first model with logistic regression with RFE which provides summary of results of the classification problem. 60.93% of true positive, i.e., the model predicted purchased and the customer purchased the device, 23.47% true negative, i.e., the model predicted not purchased, and the customer did not purchase, 9.83% false positive, i.e., the model predicted that the customer did not make the purchase, but they did, and finally, 5.77% false negative which is the model predicted that the customer would buy but they did not. The figure 12 shows the ROC (Receiver Operator Characteristic) of the model as a comparison between an ideal model and the model I created. The accuracy and precision are slightly lesser than the first logistic model with all the features, however, the recall has slightly increased, which means that the model is better than the first model at correctly identifying the positive cases. Eventually, the false negatives are also reduced from the first model.

The logistic regression model with RFE identifies “Household Income\_>100K” as the most influential factor and “Preferred color of device\_no preference” and “Household Income\_<50K” as the two highest negatively affecting factor for the prediction of “Purchased”. The R-squared is 0.426 and AIC is 5266.1339. R-squared is reduced from the first model, however, the AIC value has increased. With the increased AIC (helps analyze how well the model predicts for the new data) we can say that the second model is not as good as the first model in predicting “Purchased” attribute for the new data. Similarly, with decreased R-squared value (helps identify how well the model explains the observed values), the second model is not a good fit as compared to the first model. Hence it can be concluded that the logistic regression model with all the features performs better when compared to logistic regression model with RFE.

The last model implemented is the decision tree and Random Forest. Decision tree has a low accuracy of 78.8%, precision of 76.18% and recall of 99.3%. This algorithm combines many decision trees to effectively predict the “Purchased” attribute for the data. I split the original data into training (70%) and test data (30%). I used the training data to perform random forest which resulted in 89.07% accuracy, 91.42% precision and 92.24% recall. The Figure 13 shows the confusion matrix for the first model with logistic regression with RFE which provides summary of results of the classification problem. 61.43% of true positive, i.e., the model predicted purchased and the customer purchased the device, 27.63% true negative, i.e., the model predicted not purchased, and the customer did not purchase, 5.77% false positive, i.e., the model predicted that the customer did not make the purchase, but they did, and finally, 5.17% false negative which is the model predicted that the customer would buy but they did not. Figure 14 shows the feature importance by the random forest algorithm. Top five important features in predicting the “Purchased” attributes are “Preferred color of device\_no preference”, “Ice cream products consumed per week”, “How many desserts do you eat a week”, “ Deposit Amount”, and “Household Income\_>100K”.

Cumulative accuracy profile (CAP) for a model can be used to evaluate the robustness of the classification models. I used CAP analysis to accurately evaluate the three models above. From Figures 15(CAP for logistic regression with), 16(CAP for logistic regression with RFE) and 17(CAP for random forest). The CAP Curve tries to analyze how to effectively identify all data points of a given class using minimum number of tries. In this dataset, I tried analyzing how quickly the model can predict the “Purchased” attribute correctly. In comparison, the three above mentioned model performance shows that Random Forest algorithm is the best predictive model for the given data.

# Conclusion and Recommendations

The report helps identify some important aspects of increasing sales of “IceCubed”. The initial exploratory data analysis helped in identifying some of the customer behavior that gives insight into targeting the right customer for sales. Chocolate flavor seems to be the most favorite of all flavors for the buyers. Discounts or advertisements on devices producing this chocolate flavor can improve the sales of the devices. Targeting customers with household income of >50K helps increase the chances of device sales. From the analysis it can also be concluded that events like the campaign helps introduce the device to the customers. If correctly advertised during the campaign, the conversion rate of donors to customers is high. Predictive models for the purchase of the device were created. Logistic regression, logistic regression with recursive feature selects and random forest algorithms were implemented. Random Forest algorithm was found to be the most accurate method of predicting the purchase. Through the data modelling, it was found that if people select no preference of device color, there is a higher negative effect on converting them from donors to buyers. Similarly, if higher the number of ice cream or dessert products a customer consumes per week, higher are the chances of them purchasing the device. These insights would help Kickstarter to design a focused business strategy based on the customer behavior details mentioned above.

# References

Bhanot, K. (2019, March 10). Machine Learning Classifier evaluation using ROC and CAP Curves. Medium. <https://towardsdatascience.com/machine-learning-classifier-evaluation-using-roc-and-cap-curves-7db60fe6b716>

Huilgol, P. (2019, August 24). Accuracy vs. F1-Score. Medium. <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2>

Li, S. (2017, September 29). Building A Logistic Regression in Python, Step by Step. Medium. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

Plotly Python Graphing Library. (n.d.). Plotly: The front end for ML and data science models. <https://plotly.com/python/>

Yiu, T. (2019, June 12). Understanding Random Forest. Medium. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.>

# Appendix

Figure 1: Bar plot of categorical values in the dataset

Chart, bar chart

Description automatically generated

Figure 2: Box plot of numerical values of the dataset

Graphical user interface

Description automatically generated with medium confidence

Figure 3: Grouped bar chart of categorical values

Chart, bar chart

Description automatically generated

Figure 4: Histogram of numerical attributes of the dataset

Chart, histogram

Description automatically generated

Figure 5: Ice cream products consumed versus purchased attribues

Chart

Description automatically generated

Figure 6: Dessert consumed per week versus purchased

Chart, histogram

Description automatically generated

Figure 8: Correlation plot for variables after on-hot encoding

Chart, bar chart, histogram

Description automatically generated

Figure 9: Confusion matrix for logistic regression Chart, line chart

Description automatically generated Figure 10: ROC for logistic regression model

Chart

Description automatically generated

Chart, treemap chart

Description automatically generatedChart, line chart

Description automatically generatedFigure 11: Confusion matrix of logit reg w RFE Figure 12: ROC for logistic regression with RFE

Chart

Description automatically generatedFigure 13: Confusion matrix for Random Forest algorithmChart

Description automatically generated and decision tree

Figure 14: Feature importance for Random Forest algorithm

Chart, bar chart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedFigure 15: CAP for logistic regression Figure 16: CAP for logistic regression with RFE

Figure 17: CAP for random forest