**ALY 6040: Assignment-2 Logistic Regression Project**

Logo, company name

Description automatically generated

Spring 2022

Instructor: [Justin Grosz](https://northeastern.instructure.com/courses/105794/users/24150)

Apr 24th ,2022

Hang Wu

# Problem Background

It’s the data gathered from the Kickstarter event that happened between 2019/07/01 to 2019/07/09. There are 10000 data points in the dataset and categorized by 12 features. 5 of them are numeric features and the other 7 are category features. The data set is split into 2 samples, train and test which 80% of the data is train and 20% is test.

## Business Problems

What can company learns from the customers’ preferences and actions?

That flavor and color don’t matter as much (see analysis below), but people who owned a Keurig before will definitely buy another one.

What can company help to improve their marketing strategy?

What are the best selling products?

Nothing, in fact the lack of distinction indicates that we should worry about our client’s profile more than their preference of color and flavor.

## Data Cleaning

No data are marked as NA ,empty or Null. The kurtosis and skewness are normal (Skewness <1 and Kurtosis <3). No further cleaning is required.Table

Description automatically generated

Data Cleaning Addition to the Previous Report

This section is added to indicate the additional care done for the data.

# Model Building

The following steps should be done based on the template (*Microsoft, 2022*)

The target is set to be Purchased, every single business questions are related to purchases alone.

Step 1. Univariate analysis to find covariate/independent variable including contingency table, feature selection test and ranking test. All predictors have to be independent of each other and affecting the target only,

Step 2. Fit Multiple Logistic Modelling using the variable above. Link Function = Logit, Model = Binomial

Step 3. Check the logit plot against the covariant and check for the linearity or using the main effect plot to check how the target variable be affected by predictors.

Step 4. Effect plot being used to check the interaction effect between the target and the major predictors.

Find Collinearity: check for standard errors and use cor2cov function on the model fit found. The Alias function also provides complete collinearity analysis between variables.

Performance Testing: either use p-value or use Akaike Information Criterion (AIC) for model fit analysis, lower AIC indicates better model fit.

Feature selection are done through caret package with lvq algorithm. After the model development, model score, AIC, BIC and ROC will be used to evaluate the model effectiveness and accuracy.

# Analysis

Feature Selection Phase: when the target is set to be Purchased, feature Importance test are conducted(Figure 1 and Figure 2) below. The idea is to use widely used algorithm to find the 5 most important features.

The methods of the feature importance test are the two popular methods listed (*BrownLee, 2019*). The first model used the Learning Vector Quantization(LVQ) algorithm to solve the binary classification problem, and in turn rank the attributes by importance. The algorithm is solely used in here for **ranking** feature importance.

The test indicates that “How many deserts eaten per week” is **the most important** against all 10 other features. “Gender” and “Ice-cream consumed per week” are **50%** less important than the “Deserts eaten per week” attribute, confirming the analysis in previous report. Correlation Matrix(Figure 3) is used to further eliminate the predictors, albeit only 1 of the 5 numeric features. The most correlated Feature is removed, which is the Donated ID. **The RFE (Recursive Feature Elimination)** method shows the root square standard error of all features in terms of predicting the target(the Purchase status). Less errors presented meaning a more accurate prediction from the model. Household Income can most accurately predict the purchase status. We will be using the top 5 attributes that both methods above shared, since we want our predictors to be more important, more accurate in terms of prediction and less sensitive(less correlation): “Preferred Color of Device”, “Gender”, ”How Many deserts do you eat a week” “Donate to Kickstarter before” and “Ice-cream Products Consumed per week”. Please notice that knowing whether our customer own a Keurig is more important than knowing their favorite flavor. In fact, we are likely to sell only to diehard Keurig fans (see analysis below)

Modelling Phase: data is split in 20/80 fashion, where 20% of the testing data and 80% of the training data are randomly sampled (no systematic or clustering sampling of any kinds). A logistic regression model with the link function = logit is used for the modeling, model score, AIC and BIC will be sued to evaluate the model performance, while Confusion Matrix will evaluate the prediction accuracy and what those rates would affect our client’s profit projection.

Additional Insights to the Previous Report after knowing the target

This section is added so the conclusion is updated accordingly as the previous report dismiss several important insights. It is definitely good to know after the feature selection process. Choice of color doesn’t really differentiate against the purchase event(Figure 5), meaning that team red isn’t more likely to buy than team blue. However if the preferred choice isn’t chosen, it’s 13.3% likely they wouldn’t purchase. Selecting a color alone gives them 10% more chance to purchase. 7 is the magic number of deserts eaten per week (Figure 6), people either eat more than that or less. People who eat less than 4 severely decrease their chance to buy, from approximately 2% to 8.3% (12.1%-3.8%), 7.3% and 8% respectively. This means that a clear threshold of 4, and 7 can be potentially used to predict the probability to purchase. Please refer to previous report’s insights that there isn’t a clear distinction of ice-cream counts (Figure 7 also indicates that), but a clear overall count distinction can be seen by the number of deserts question. In fact, if more than **70%** of the people participated in this event own a Keurig(Figure 8), you know that people don’t buy Keurig only for function purpose, **Keurig diehard** would definitely own more than one. Not only that, they are 2.5 times more likely to deposit more than $200 when the minimum is $100(1 to 5 ratio when a person who doesn’t own a Keurig, 2 to 1 ratio for person who owns a Keurig). Due to that statistic, we can predict that 17.5% of the deposit amount to be at least doubled of the original amount of $100, this would give us a fair start for initial revenue stream, and it would allow us to develop a subscription model in addition to the pre-sales. If there is one thing we know, chocolate lovers (Figure 10) definitely is 2.8% more likely to buy and 3% less likely to say no. Even though flavor is not an important feature, it’s still an interesting fact for our customer base. Also

# Interpretation and Recommendation

The first thing our client can learn that “color of the device” can statistically more accurately predict the purchase status than “favorite flavors”. In terms of producing a survey for our client, “Owning a Keurig” might seem important in common sense, but it isn’t as important in terms of predicting the results, the same with knowing the gender and surprisingly, knowing how many ice-cream consumer per week.



Why do our Odds Ratio/Exponentiated Coefficient Matter to our clients?

The Odds Ratio for ice-cream consumed per week indicates 4% (OR=1.04)more odds that ice-cream consumed matter to whether purchase our client’s Keurig, which is 10% less than the odds ratio for knowing how many deserts consumed per week (OR = 1.14). Knowing the gender actually decreases the overall odds (male OR=0.53, female OR=0.69) by at least 47%. However, knowing whether the customer is female increase the odds by 16% percent comparing to male.

Chart

Description automatically generated

What does the confusion matrix tell us about our prediction?

With FP=439,TP=1133,FN=179,TN=249 we know several things. We have a prediction accuracy of (1145+222)/2000=68.35%, which fluctuate within 5% approximately depending on how many times the simulation is run, making the model not overly sensitive. Of course, True Positive results would be slightly more important because that would indicate our expected inventory and thus the projected revenue.

24.5% (490/2000) of the units we think would be sold turns out not to be, we assume that each purchasing event limits to 1 Keurig per person. And we assume that people who donates more than the unit price (likely to be 100, the minimum of the donation) won’t get extra units. If that’s the case a false positive (FP) rate of 24.5% indicates that 24.5% of the projected revenue for our client’s company would be false, of course if we take the False Negative into account (11.1%) we would decrease our projection failure into 13.4%. Indicating that our projected inventory is overestimated by 13.4%, and total profit is overestimated as a result. Of course, the inventory fee is not necessarily linked to the number of units stored, It’s still worth noting that improving our model would save us at most 13% of the space, by the average Keurig size(*Coffee Maker Support,2022*) that equates to 20 by 30cm, we can save at most 53.60 meter by 80.40 meter of inventory space per 2000 units.

RoC indicates a 70.85% AUC (Area under curve), which is poor considering it’s closer to 0.5 than 1. AIC and BIC are 9393.486513687269 and 9456.371285073226 respectively, we do expect a good model to have less than 1000 AIC or at l east less than 2000, although changing features might change, a newer model is expected in future analysis to better classify the information.

## Quick Summary

Since our analysis indicates that diehard dominates our client demographic (people who owns Keurig but buys anyway), and we could overestimate our sold unites by at least 1/5. Plus we don’t have to worry about the color or the flavor. We can build a marketing strategy to focus on selling more units to the existed donors. Although the accuracy and AIC don’t seem to be top level, the predictors shows statistically significance (p -value<0.05) in affecting the target outcome, and there is a statistical significant distinction between purchased and non-purchased event/profile.

# Conclusion

Not only we derive the top 5 features from the 12 attributes, we find the distinctive difference in attribute that allow us to further optimize in order to get the best result. We use the Logistic Regression with the Link function to be Logit, and use a 80,20 split for training/testing dataset. The final result shows a 68.35% accuracy which is also confirmed by model score, and a fairly low rate (13.4%) of overestimation of the number of orders. ROC curve shows a AUC of 70.85%, which is not good for our model efficiency. We expect a more complex model with much better AIC/BIC, model score and ROC so we can rely on.

# Appendix

Figure 1. Feature Importance Test on all Features

Chart

Description automatically generated

Figure 2. RFE Method for Feature Selection

Chart, line chart

Description automatically generated

Chart

Description automatically generated

Figure 5: If color choice affects the purchase intent

Chart, bar chart

Description automatically generated

Figure 6: deserts eaten per week vs the decision to buy

Chart, bar chart

Description automatically generated

Figure 7: Count Plot of icecream per week

Chart, bar chart

Description automatically generated

Figure 8: Are our clients mostly Keurig diehard who owns more than 1 Keurig? Yes. Overwhelmingly

Chart, bar chart

Description automatically generated

Figure 9: Do people who own a Keurig more likely to deposit more than $200?

Chart, bar chart

Description automatically generated

Figure 10: does chocolate lovers more likely to buy? Yes, they do by 2.8% and 3% less likely to pass on the opportunity

Chart, bar chart

Description automatically generated

Figure 11: Model Results

Table

Description automatically generated

Figure 12: ROC

Chart, line chart

Description automatically generated

Figure 13: MSE and R2

Chart, bar chart, treemap chart

Description automatically generated

# Reference

Sally Bitterl and Martin Schreier (2016) ,"More Than Just the Money: Psychological and Behavioral Consequences of Participating in Crowdfunding", in NA - Advances in Consumer Research Volume 44, eds. Page Moreau and Stefano Puntoni, Duluth, MN : Association for Consumer Research, Pages: 394-395.

Building and Applying Logistic Regression Models, Microsoft, 2022, <https://personal.utdallas.edu/~pkc022000/6390/SP06/NOTES/Logistic_Regression_4.pdf>

Brownlee, J. (2019, August 22). *Feature selection with the Caret R Package*. Machine Learning Mastery. Retrieved April 22, 2022, from <https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/>

*Coffee Makers Support*. Keurig® Coffee Makers FAQs | Keurig®. (n.d.). Retrieved April 24, 2022, from https://www.keurig.ca/support/keurig2-brewers