Data Science for Business: Team Project Final Report (iFood Case)

Team 30 – Section B

Jialin Hu, Shuying Wang, Mahima Kriti, Mandy Zhou, Haoming Bai

---

BUSINESS UNDERSTANDING

iFood is an online food and grocery ordering and delivery company based in Brazil. In the past three years, the company has had solid revenue and a healthy business structure. However, the growth trajectory for the next three years is not as promising. Therefore, iFood hopes to reverse the declining growth trend and make it positive through its marketing campaigns. Our goal is to determine the strategy for the next marketing campaign that can generate the highest profit for iFood. To achieve this, the company aims to: **Efficiently exclude non-respondents from the campaign to minimize wasted resources. Understand the characteristics of customer segments more likely to accept the ad campaign.**

Data mining : The data mining tasks we wish to perform are as follows:

- Customer Segmentation: Data mining will analyze past campaign data to identify customer segments based on demographics, behavior, and purchase history. This will help in targeting the right customers and tailoring marketing efforts to specific segments, increasing the likelihood of gadget purchase.

- Predictive Modeling: By building predictive models, data mining can identify customers who are most likely to purchase the gadget. These models can assign a probability score to each customer, allowing the company to focus marketing efforts on high-probability customers, thereby maximizing campaign profitability.

- Efficiency Improvement: Data mining will assist in creating efficient exclusion criteria for non-respondent customers. By identifying characteristics of non-respondents, the company can save resources by not targeting those who are less likely to engage with the campaign.

DATA UNDERSTANDING

- Data Overview:
  - The dataset has 2240 observations (representing 2240 customers).
  - There are 25 features in the dataset, including one target variable, "Response."
- Features:
  - Demographic Information: gender, year of birth, education, marital status
  - Purchase Patterns: Columns like "MntWines," "MntFruits," "MntMeatProducts," etc., provide information on the amount spent on various product categories and the number of purchases through different channels.
  - Response to Previous Campaigns: Columns "AcceptedCmp1" to "AcceptedCmp5" indicates customer responses to previous campaigns.

DATA PREPARATION

- Education: There are 6 different categorical columns converted to dummy variables
- Marital Status: The original dataset had 6 distinct categories of marital status we converted them to three categories: Single, Married and Together

- New fields created: Age, Days_until, MntTotal

- Imputation: Income had 24 NA values, they were imputed by simple linear regression

- Dropping columns: Z_Revenue(column of 3s), Z_Cost, Dt_Customer, ID

Final dataset:

- 2240 observations and 32 fields

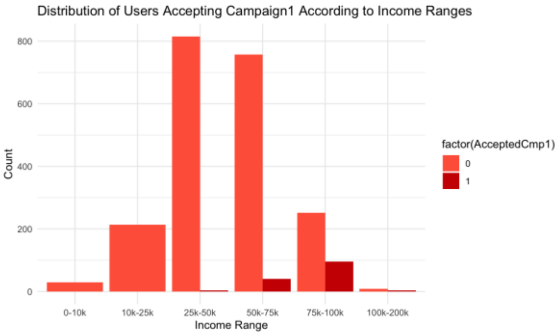
DATA EXPLORATIONS

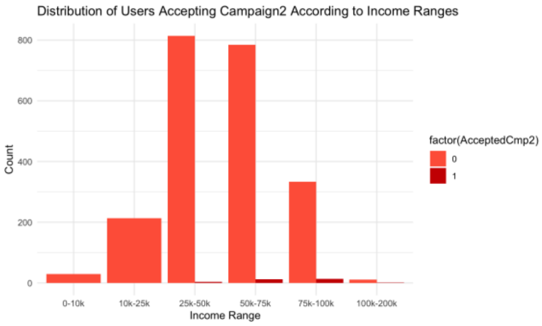
- Our consumer group earns around 50,000, and it comprises more highly educated individuals compared to those with lower educational levels. *(check graph in Appendix- 1a)*

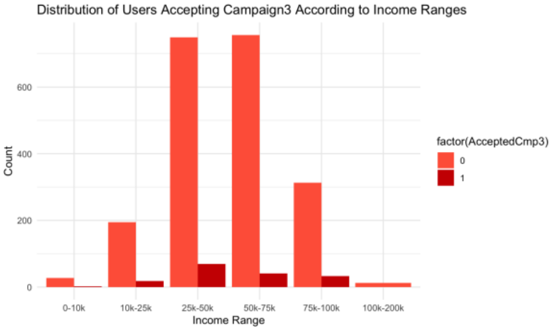
- July has the lowest number of enrollments, indicating a challenge for customer acquisition. To optimize the profit, we should avoid and launch the next campaign during other peak seasons. *(check graph in Appendix- 1b)*

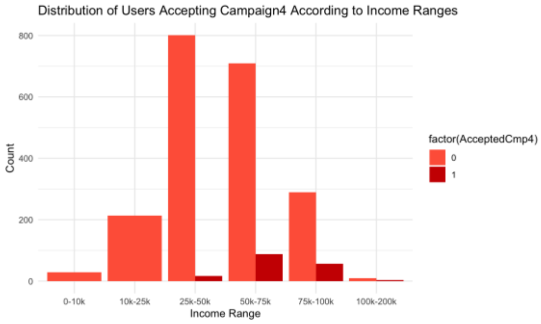
- **<u>How are various income groups responding to the various campaigns?</u>**

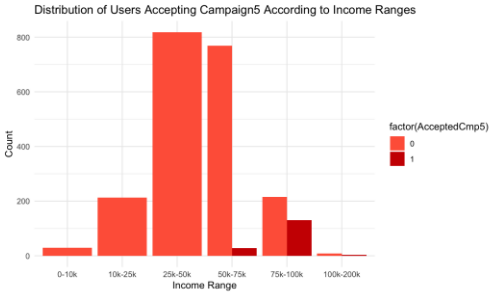
From the result, we can see that the income group of 75-100k responded the best to Campaign 5, income range 50-75k responded the best to campaign 4, and income range 25-50k responded the best to campaign 3. Therefore, when we design the next campaign targeting different income groups, we should base it on the trend we see in the past data. **Another noticeable trend is that people with higher income groups respond better or more easily to our campaigns.**


Distribution of Users Accepting Campaign1 According to Income Ranges


Distribution of Users Accepting Campaign2 According to Income Ranges

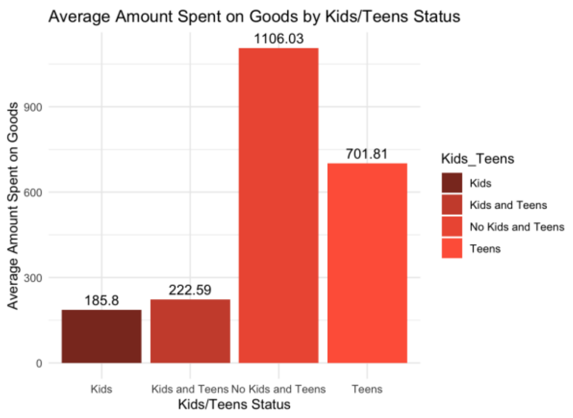

Distribution of Users Accepting Campaign3 According to Income Ranges

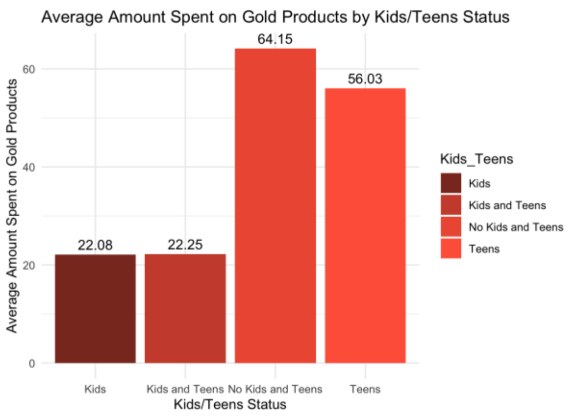

Distribution of Users Accepting Campaign4 According to Income Ranges

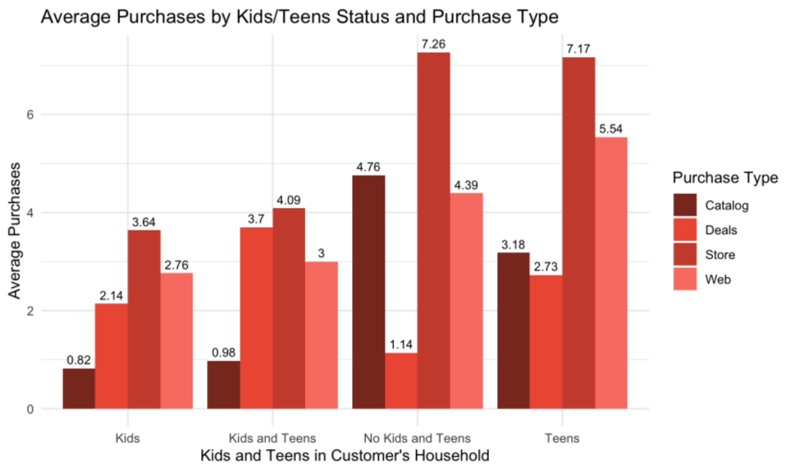

Distribution of Users Accepting Campaign5 According to Income Ranges

- **<u>How are households with/without kids/teens purchasing different products</u>**

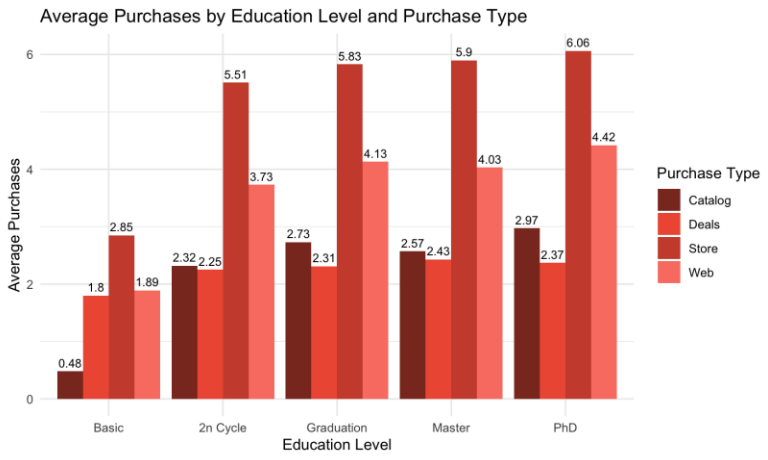
Houses without kids or teens spend the most on both Regular and Gold products and they use all 3 sales channels – Store, Online and Catalogue.





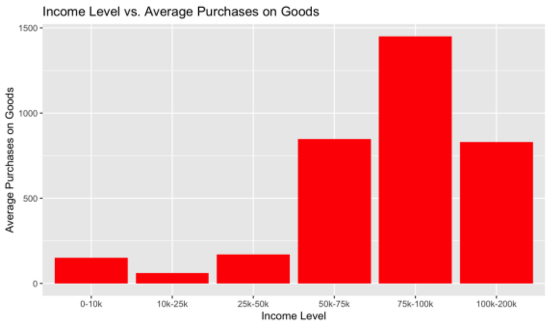


- **How does education level affect purchasing habits?**

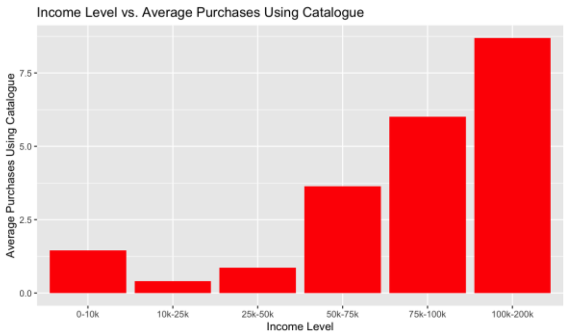


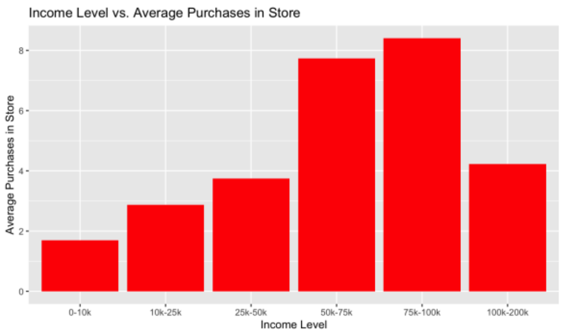
- **How does income type affect purchase patterns?**

Higher income groups (75k and above) spend more on goods compared to lower income

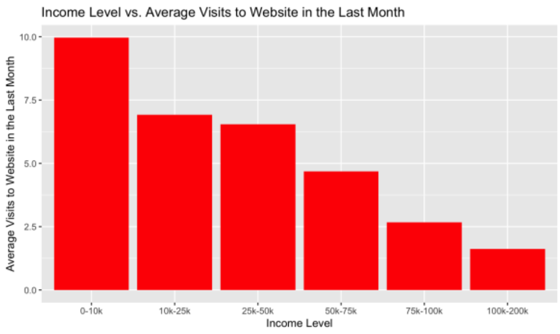


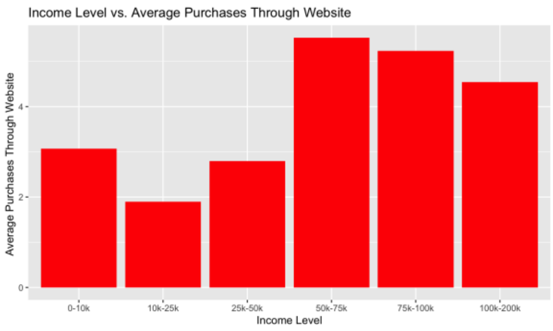
Higher income groups purchase significantly via both catalogues and in-store





- People with higher income make more web purchases but the lower income groups visit the iFood much more compared to higher income groups.





These six graphs show the relationship between Income level and other factors. We divide Income level into 6 levels, from 0-10k to 100k-200k.

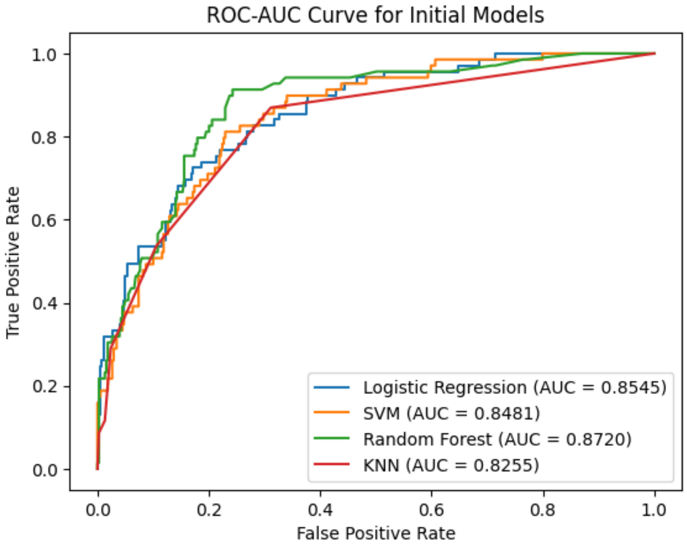
Recommendation : On the basis of our data exploration iFood should target high income category consumers who have an advanced degree of education and have no kids or teens .

MODELLING

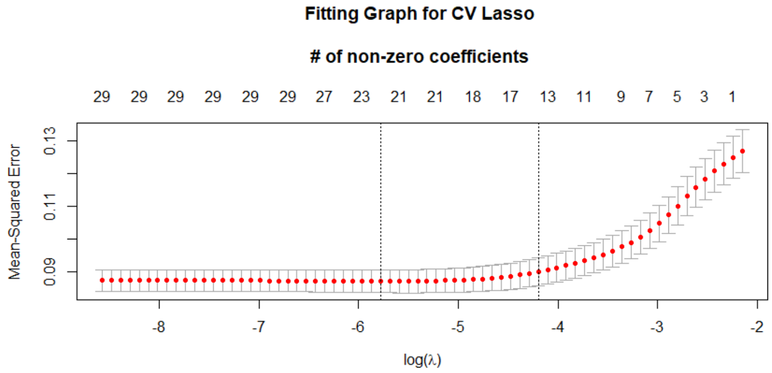
We first identify this as a classification problem.

Models - logistic regression, random forest, SVM and KNN.

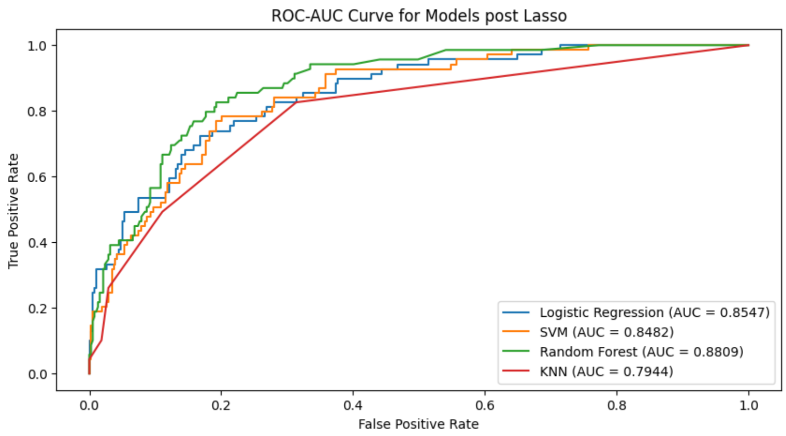
ROC-AUC curve analysis – To choose our threshold we used ROC-AUC curve analysis to find the threshold which has the highest sensitivity/recall.



We then tried Lasso as our regularization technique to select our features and it choose 22 features



Post choosing features using Lasso we again tried the 4 different models and measured the AUC score again which was highest for Random Forest at a **threshold of 0.18** as we see from the graph below.



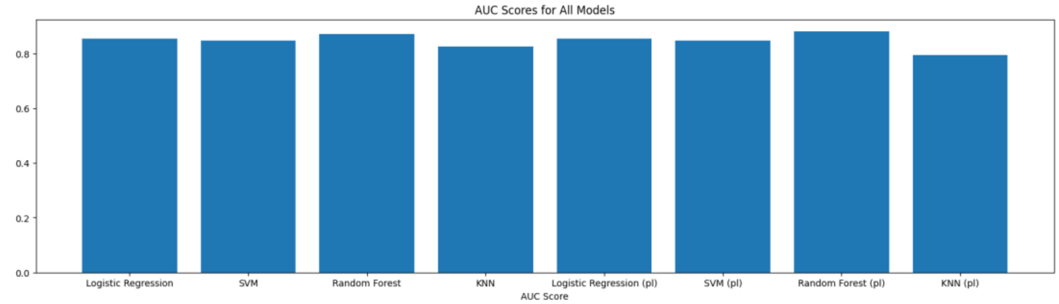
EVALUATION

For this business problem we have used AUC as our evaluation metric. The AUC curve helps us understand our model performance across all thresholds and find the optimum threshold by balancing the True positive Rate and False Positive Rate.

- We want to focus on people whom our model said would respond to our campaign and they did and We also want to focus on those people who our model said would not, but they responded “Yes” to our campaign – iFood can neglect spending marketing efforts on these set of people because they are going to accept our campaign

- Secondly, we want to direct our marketing resources towards consumers who our model predicts will not respond to our campaign

By focusing on AUC, iFood will have a balanced strategy towards marketing by avoiding spending money on that segment of its customer base that will respond favorably to its campaign and instead, they will be able to direct its budget towards other segments of the population.



Deployment

- **iFood can feed its customer data into our model and** forecast whether iFood’s prospective customers will respond positively to a campaign.
- The predictive model we constructed will improve campaign performance and targeting accuracy by identifying potential consumers who are likely to engage with the upcoming campaign early on, thus optimizing marketing resources.

Deployment – Risks and Future Considerations

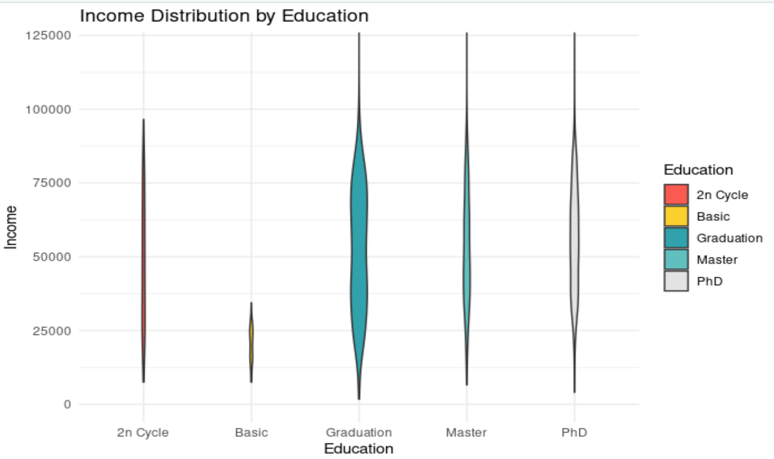
- The dataset (2240 observations used for prediction may not be representative of iFood’s entire consumer base hence our model may not perform well on population data.

- Dataset is highly imbalanced, only 15% of the customers responded “Yes” to the 6th marketing campaign.

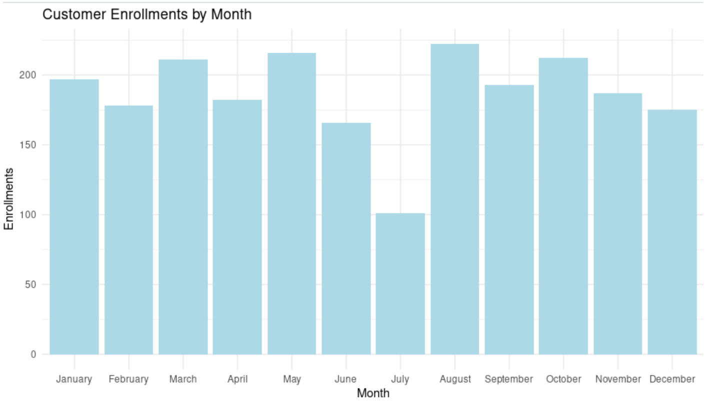
- Uncertainty with Customer Habits: Customer behaviors may change over time during big events such as the pandemic, getting married, or re-location. This model is effective at the current time, but overall, iFood should consider updating its database consistently to ensure that they are up to date in understanding their customer behavior and to effectively apply predictive analysis.

- Data Privacy Issue(ethical consideration): As the customer data is collected from the company’s customer management system, iFood should refrain from sharing any personal information with the public. When running a campaign with a third-party marketing consulting agency, iFood should restrict data usage solely for prediction purposes.

Appendix


Income Distribution by Education
Income
125000
100000
75000
50000
25000
0
2n Cycle
Basic
Graduation
Master
PhD
Education
Education
🟥 2n Cycle
🟨 Basic
🟦 Graduation
🟩 Master
⬜ PhD

1a


Customer Enrollments by Month
Enrollments
200
150
100
50
0
January
February
March
April
May
June
July
August
September
October
November
December
Month

1b