Bookbinder Study Case

Alex Martinez, Josh Gardner, Cameron Playle, and Guillermo Gallardo

2024-09-25

Paper Starts Here

Executive Summary

Brief introduction of problem. Summarizes key findings. Summarizes insights behind key findings.

Our Problem

Clear description of the problem, from an application and theoretical point of view. Outlines the report.

For this study case our goal is to evaluate how effective three different models are and compare them with the option of creating this campaign without a model. We are trying to determine which model will provide the best balance between cost savings and profit. By analyzing the performance of each model and comparing ti against the campaign, we will identify the most cost-effective approach that maximizes profit.

The three models we are comparing are the Linear Model (LM), Generalized Linear Model (GLM), and Support Vector Machine (SVM). Although we anticipate that the linear model may not perform well, we are still interested in understanding why it may not be the best fit for this case study. This exploration will help us gain valuable insights into the limitations of the linear model in this context and guide our decision-making process.

Literature Review

Discusses and cites existing works in the theoretical and application realm.

Methods

Discusses types of variables, sample size, and sampling techniques (if any). Discusses the model(s) and its assumptions and limitations.

Our dataset was given to us divided into training and test sets. The training set includes 1,600 observations, while the test set contains 2,300 observations. The dataset consists of 12 variables, with one variable (observation) being removed. Two variables were converted into factors: gender and choice, with choice serving as our response variable. The remaining variables are numerical. For some of these numerical variables, including Last_purchased, we will transform them into categorical variables to see how it would impact our modeling process.

Add stuff about unabalanced dataset?

GLM:

SVM: balance vs unbalanced comparison?

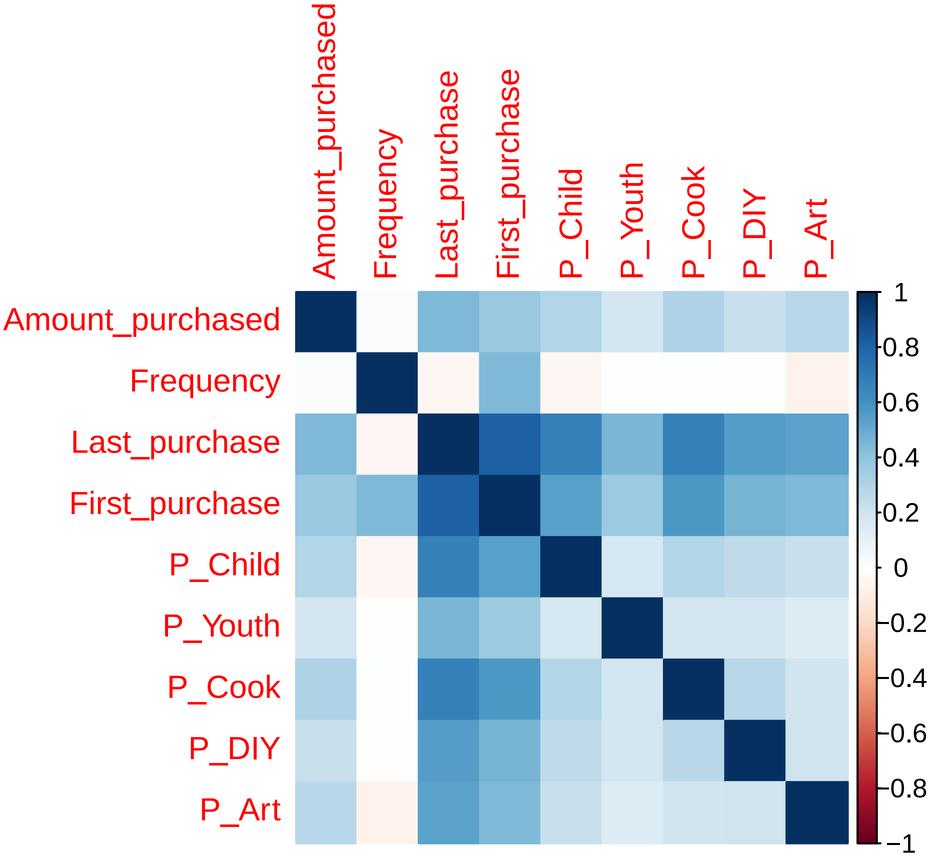
LDA maybe?

Data

Discusses how data was handled, i.e. cleaned and preprocessed. Discusses distributions, correlations, etc.

Correlation

The graph below highlights the variables with the highest correlations. We observed that *first_purchased* and *last_purchased* exhibit the strongest correlation. Based on this, we decided to create labels for these two variables to test their potential impact on our model's performance.. **ADD STUFF ABOUT RESULTS. DID IT WORK OR DID IT MAKE IT WORST?**



Results

Presents and discusses the results from model(s). Discusses relationships between covariates and response, if possible, and provides deep insights behind relationships in the context of the application.

Conclusion

Concludes with a summary of the aim and results. Discusses alternative methods that can be used.