94-881: Managing Analytics Projects (Spring 2024)

Final Project Presentation and Report

Superstore Marketing Campaign

Jingchen Fu (jingchef@andrew.cmu.edu, jingchef)
Jiuyuan Xie(jiuyuanx@andrew.cmu.edu, jiuyuanx)
Minkang Li(minkangl@andrew.cmu.edu, minkangl)
Yi Liu(yiliu2@andrew.cmu.edu, yiliu2)

Prof. David Steier

Carnegie Mellon University
Heinz College

April 30th, 2024

1. Executive Summary

Background

Superstore is initiating an end-of-year sale offering a discounted Gold membership, aiming to enhance customer loyalty and increase sales. To ensure the success of this campaign, Superstore plans to use a predictive model to identify customers most likely to purchase the membership, thus optimizing marketing efforts and reducing costs.

Strategies and Implementation

- **Data Utilization:** Leveraged data from a pilot campaign, including demographics and purchase behaviors, to train classification models (Logistic Regression and XGBoost).
- **Model Development:** Addressed challenges such as data imbalances and variable scaling to enhance model performance.
- **Predictive Analysis:** Utilized SHAP values and other statistical measures to understand feature importance and guide the targeting strategy.

Results

The models demonstrated a strong ability to predict customer responses, with the XGBoost model showing superior performance in terms of precision and recall balance. Key insights from segmentation indicated that customers with low recency, high and medium visiting frequency, and high spend in the store were more likely to respond positively.

Future Recommendations

- **Model Refinement:** Further refine the models by eliminating non-essential features and improving their generalization capabilities to enhance predictive accuracy.
- **Customer Insights:** Deepen the analysis of customer behavior through additional data collection and causal inference studies to better understand the drivers of membership purchases.
- **Technology Integration:** Ensure seamless integration of the predictive model with existing CRM systems for efficient campaign execution.
- **Stakeholder Engagement:** Maintain robust communication with all stakeholders through regular updates, meetings, and real-time dashboards to ensure alignment and adapt strategies as needed.

Conclusion

The application of predictive modeling in Superstore’s marketing strategy has set the stage for a more targeted and efficient campaign, promising not only higher conversion rates but also enhanced customer satisfaction. Continued improvements in model accuracy and integration with business operations are expected to further capitalize on these gains.

2. Problem Framing

2.1 Organizational Context

Superstore, a local supermarket, is planning an end-of-year sale targeting its existing customer base with a special offer: a discounted Gold membership that provides 20% off all purchases. Normally priced at $999, the Gold membership will be offered for $499 during this promotional period. To optimize the outreach of this campaign, Superstore intends to use a predictive model to determine which customers are most likely to purchase this membership, thereby reducing the overall costs associated with the phone call campaign that is planned to market this offer.

Decision to be improved: The decision being improved is the selection of customers to contact via phone calls. The goal is to increase the efficiency of the campaign by focusing on those customers who are most likely to respond positively to the offer, thereby reducing unnecessary marketing expenditures and improving the success rate of the campaign.

Decision Maker: The marketing team at Superstore, supported by the data science team, is responsible for making decisions based on the predictive model's classifications. The marketing team will use these insights to strategize the phone call campaign and ultimately decide which customers to contact.

Value of the Improved Decision: An improved decision will lead to a higher conversion rate, meaning more memberships sold with fewer phone calls made. This efficiency gain not only reduces direct marketing costs but also enhances customer experience by targeting only those likely interested in the offer, potentially increasing customer satisfaction and loyalty.

Baseline of the Decision: According to the existing labeled data, among all customers, the proportion of customers who purchased the Gold membership accounts for 14.9%. This is our baseline when we did not adopt the model. We hope to increase the conversion rate to 40% after adopting the model.

2.2 Requirements and Assumptions

Business Requirements:

- Maximize the response rate of the campaign by targeting the right customers.
- Reduce marketing costs by minimizing the number of unsuccessful phone calls.
- Maintain or enhance customer satisfaction and loyalty through personalized and relevant offers.

Technical Requirements:

- Develop a predictive model that can accurately identify customers likely to purchase the Gold membership.
- Ensure the model integrates seamlessly with existing CRM systems for efficient implementation of the campaign.
- Utilize data analytics to continually assess and refine the model's performance.

Assumptions:

- **Project Context:** The offer is appealing and priced competitively enough to motivate a significant number of existing customers.
- **Staff:** Assumes the presence of a skilled data science team capable of developing and deploying a predictive model, as well as a marketing team experienced in managing targeted campaigns.
- **Technology:** Assumes that Superstore has the necessary IT infrastructure to support data integration, model development, and deployment, including a CRM system that can be used to manage the campaign.
- **Data:** Assumes access to comprehensive and clean historical data on customer purchases, demographics, and previous campaign responses, which are essential for training the predictive model.

3. Initial choices for Version 1

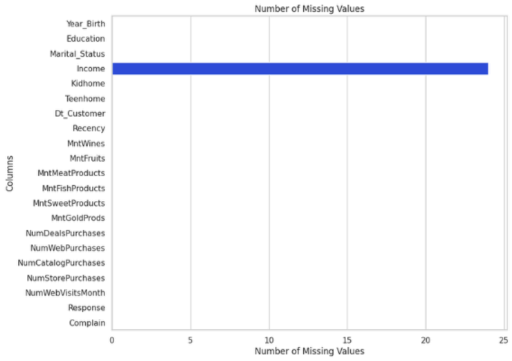
Data source:

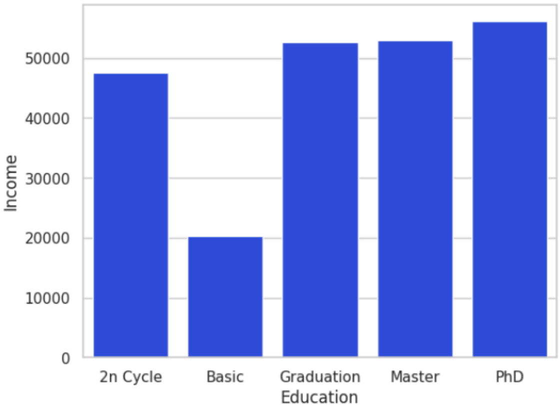
The data comes from a pilot marketing campaign of the superstore on 3000+ customers, with their demographics(age, marital status, education, income…) and behaviors(total spend, frequency of visits, recency…) information and with the target variable response(whether they responded to the marketing campaign), and we build classification model(Logistic Regression and XGBoost) for classifying whether people are going to respond or not to our campaign.

Data Preparation:

1.Null Value:

Null Values

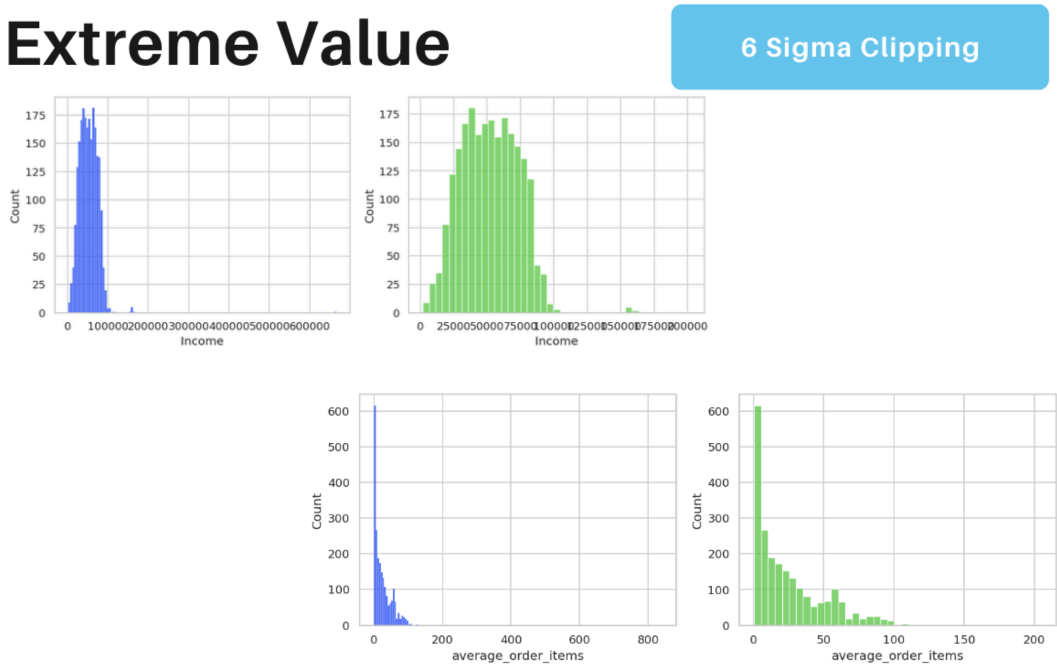




Fill with mean income in the education group

Our data only has null value in the income column. Also we find that income is highly correlated with education level, therefore we fill in null values based on mean income of the people who have the same education level as the person. This method may introduce bias to our model, which is a trade off between less data points with no mistaken information by dropping them directly and more data points with information that may not be accurate by filling with mean.

2. Extreme Value

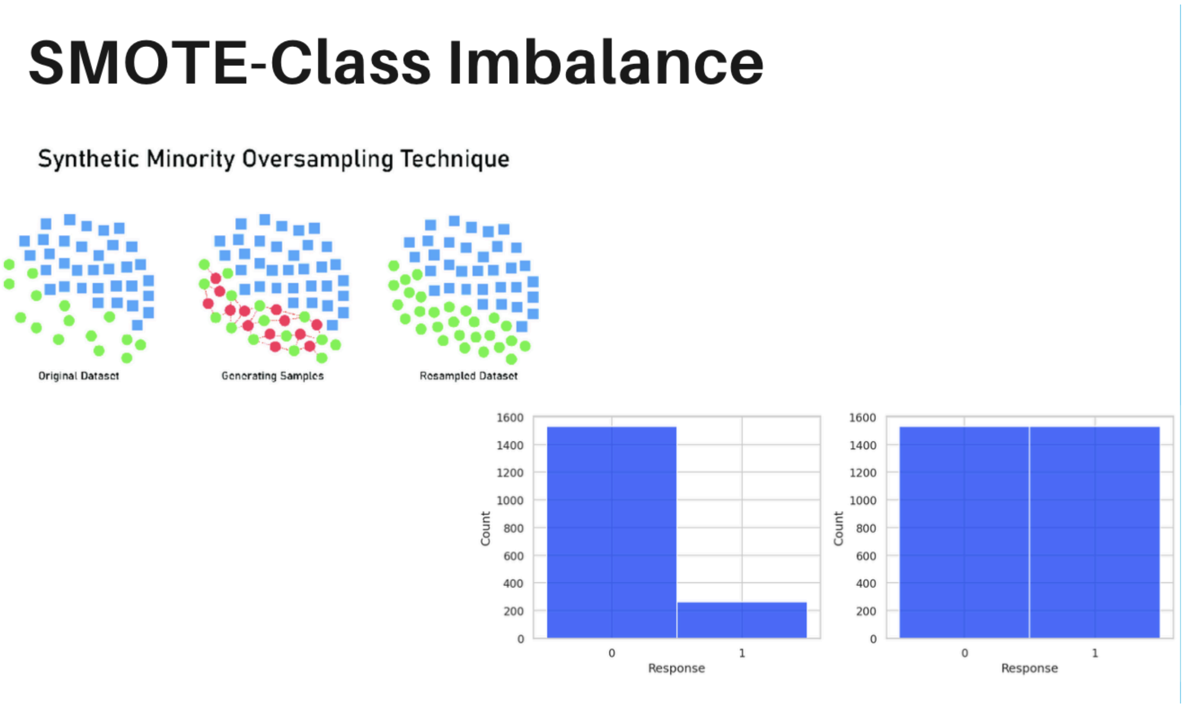


Extreme Value

6 Sigma Clipping

We can see from the graph income and average items per order contains extreme values that could affect the model's performance negatively, then we used 6 sigma clipping to clip the value to be within 6 standard deviations away from the mean.

3.Class resampling



SMOTE-Class Imbalance

Synthetic Minority Oversampling Technique

We find that our responded people only consist of 14.9% of the total population, which is a significant class imbalance problem, which can affect our model's performance negatively, therefore we perform SMOTE upsampling methods to balance the class.

4. Standardization:

Our data distribution varies a lot, in order for logistic regression to perform well, we need to use Standard Scaler in sklearn(subtract the mean and divide by standard deviation) to make each feature's distribution to have zero mean and unit standard deviation.

5.One hot encoding:

The feature education and marital status contains categorical values, but we don't want to integrate essential ordering into the values, therefore we used one hot encoding (pd.get_dummies) to convert the feature into numerical values so that it can work in machine learning models.

Which data visualization techniques best communicate the necessary information for each stakeholder?

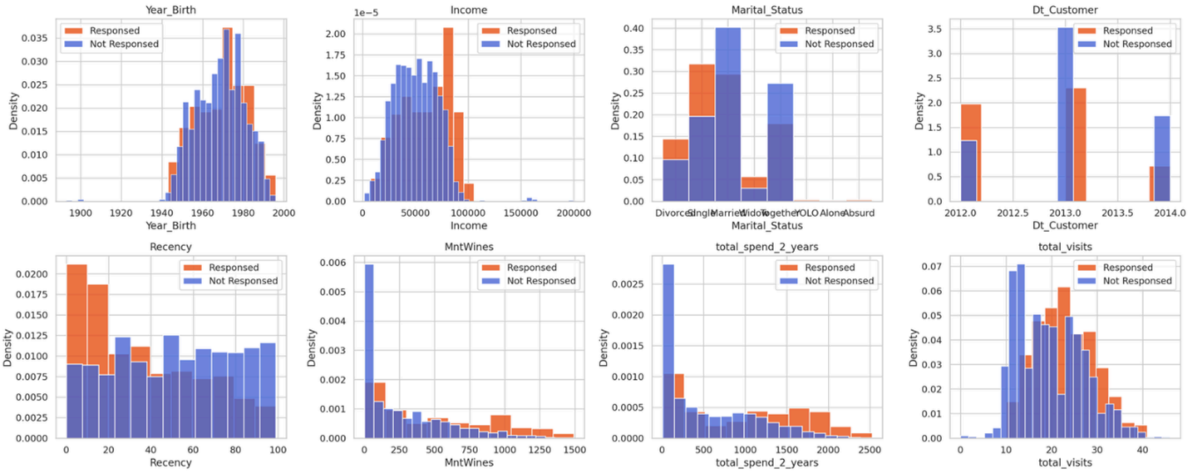
Overall we think the XGBoost's SHAP value is the best visualization to communicate the information, because it helps us understand each feature's importance in the model's prediction of response, and helps the marketing team understand and precisely target people.

Also with our model's AUC and ROC we can show the credibility of our model.

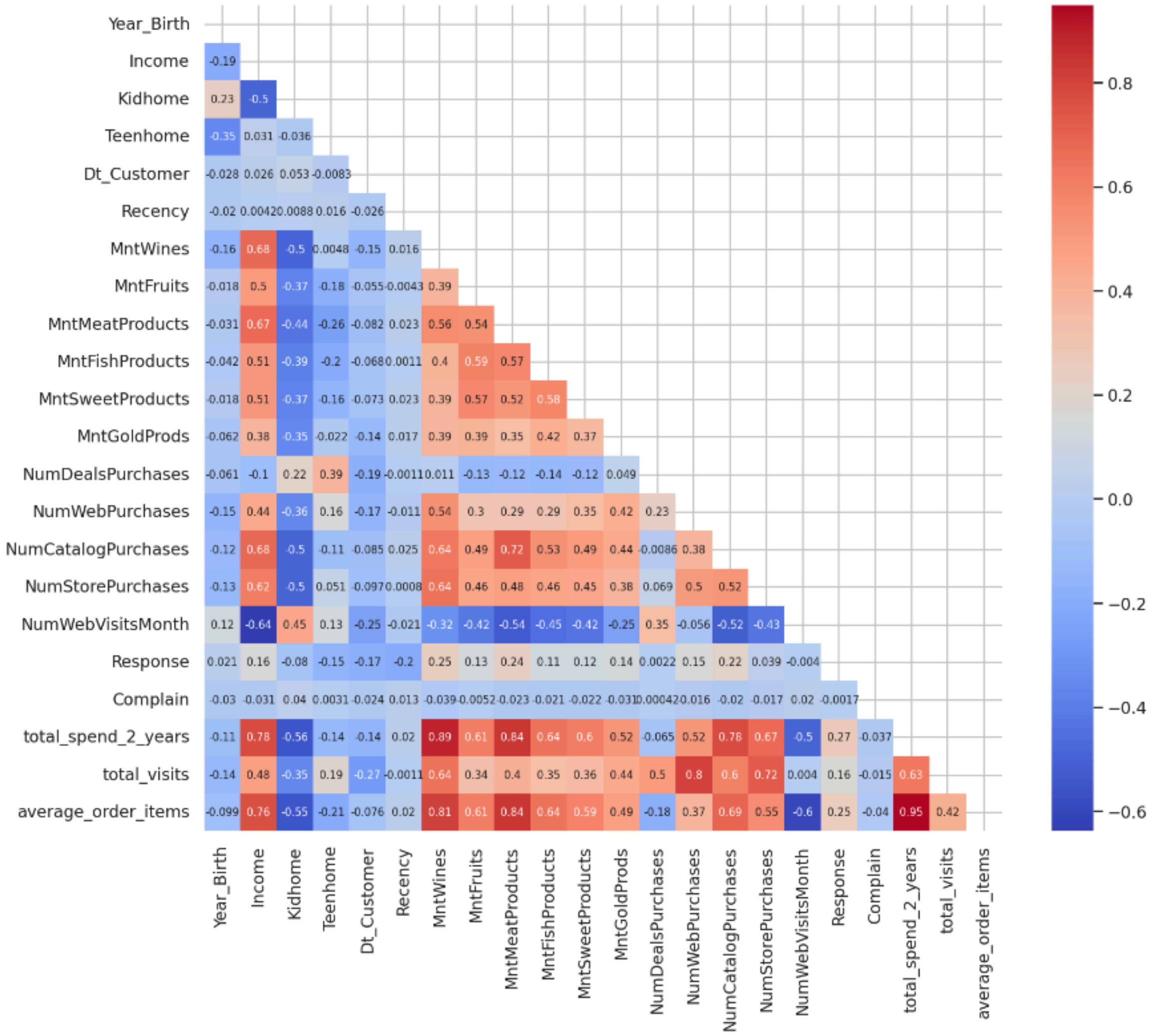
Here are all the plots we find necessary:

- Bivariate analysis

Use response=0/1 as the grouping standard, and use histogram to display the group distribution of each variable, and overserve key demographic or behavioral differences in people who respond to the offer versus those who don't.

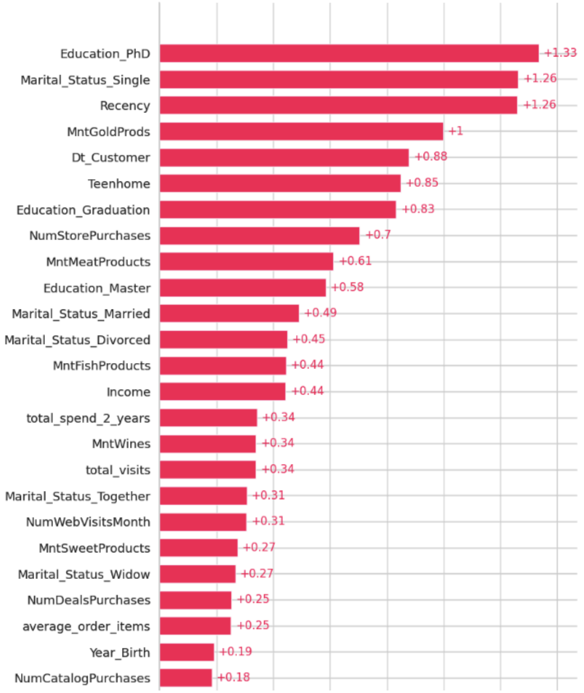


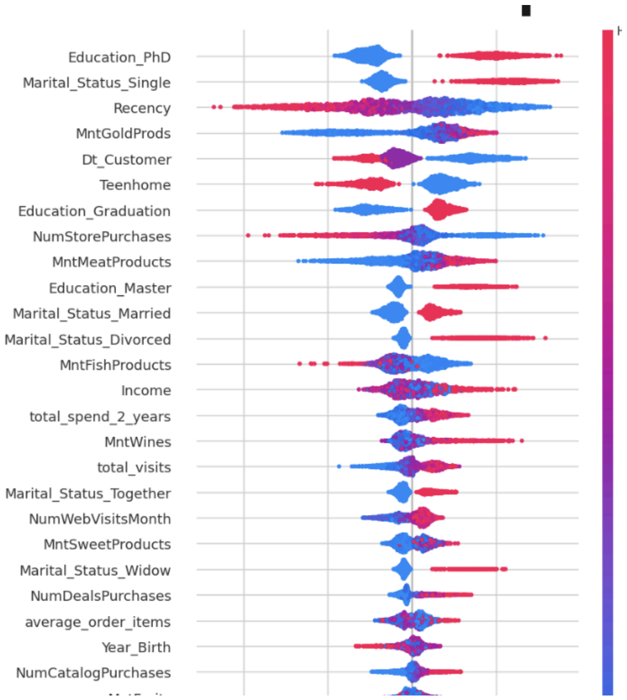
Use heatmaps to show correlations between variables. Particularly observing the correlation of Response with other features, we can see that response is correlated with our features. And there are multicollinearity problems, which guides feature selection.



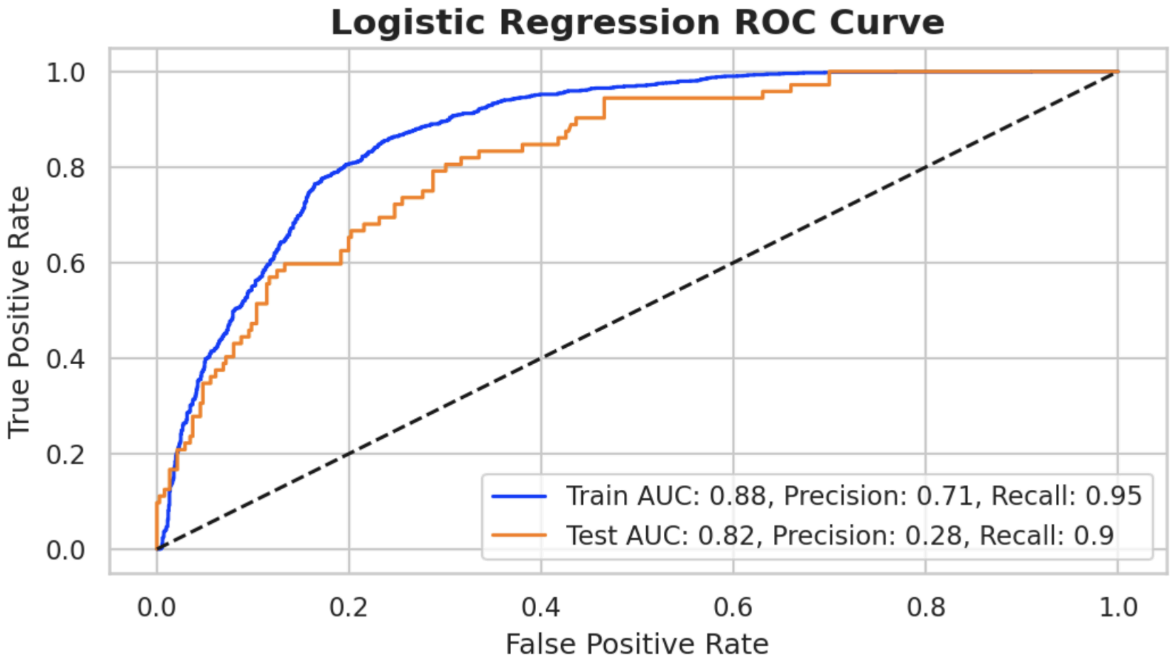
- Key findings from the predictive model

Display the features in the XGBoost model and their impact on the prediction results through beeswarm plot. The impact is measured in SHAP values. Pink value means the feature is large, blue means the feature is small. And if pink appears on the right side, it means the large value makes the prediction towards positive class(respond to campaign) versus on the left side which means towards negative class. And it is sorted from the most important features to the least important features. Here we can see that people with Education_PHD, Marital_Status_Single and low recency are more likely to respond positively to our offer, and we should target them in higher priority.

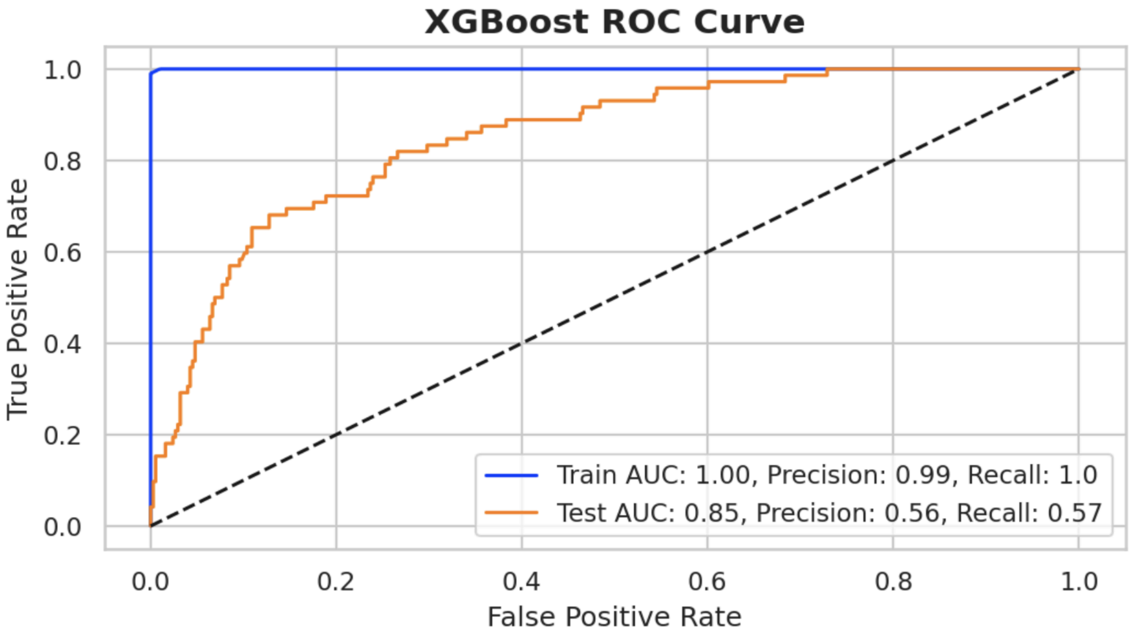




- Model performance metrics

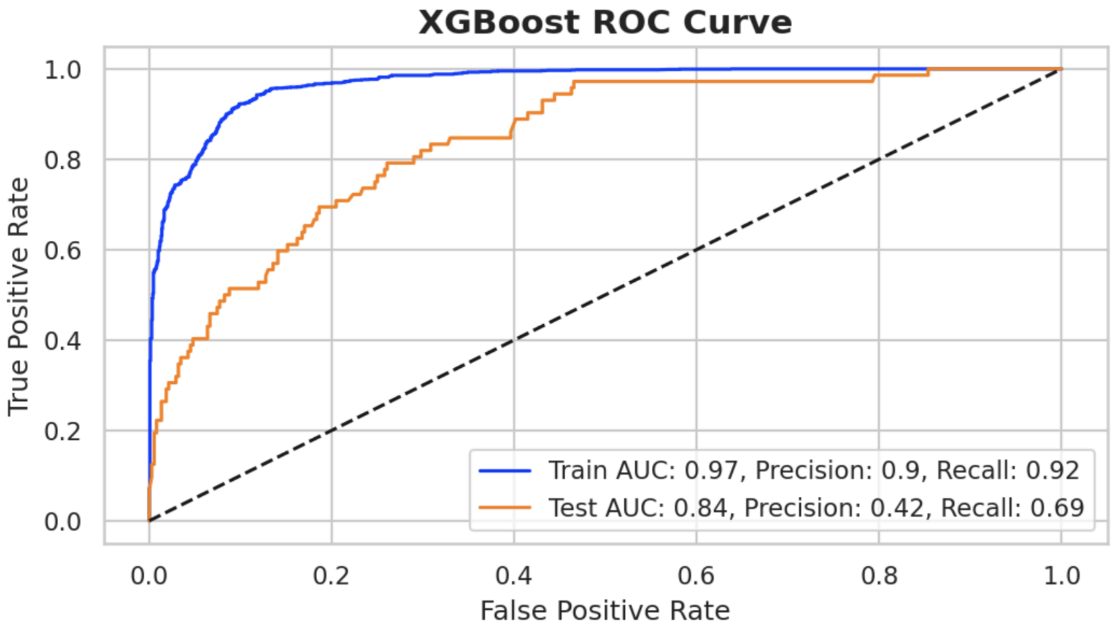


Display the ROC curve and AUC to show the model performance. ROC curve is an important measure of classification model's tradeoff of precision and recall. It shows the model's performance under all classification thresholds. A higher Area Under the ROC Curve(AUC) means the model has better and more robust performance under classification thresholds, and helps us decide the best threshold. Here our Logistic Regression model achieves a test AUC of 0.82, and 0.88 for XGBoost which is a nice performance of classification model.



The train ROC of the XGBoost shows perfect classification which can be due to the tree model splits forever until each decision node contains one training data point, which is a sign of overfitting and the evaluation performance will be low. It can be mitigated by reducing the number of trees in the ensemble model, and reducing the depth of trees, and setting the smallest number of children in a node, lowering the learning rate and increasing regularization term in the XGBoost model.

After we adjust the parameters, we didn't get a higher test AUC, but we got a higher recall, and lower precision:



Reasons for choosing Logistic Regression and XGBoost:

Logistic regression is a simple linear binary classification model, which is highly interpretable, which serves as a baseline model.

XGBoost model is an ensemble gradient boosting decision tree model that can capture nonlinear relationships between features, and feature interactions through splitting nodes, and is generally very high performance and fast inference speed due to parallelization.

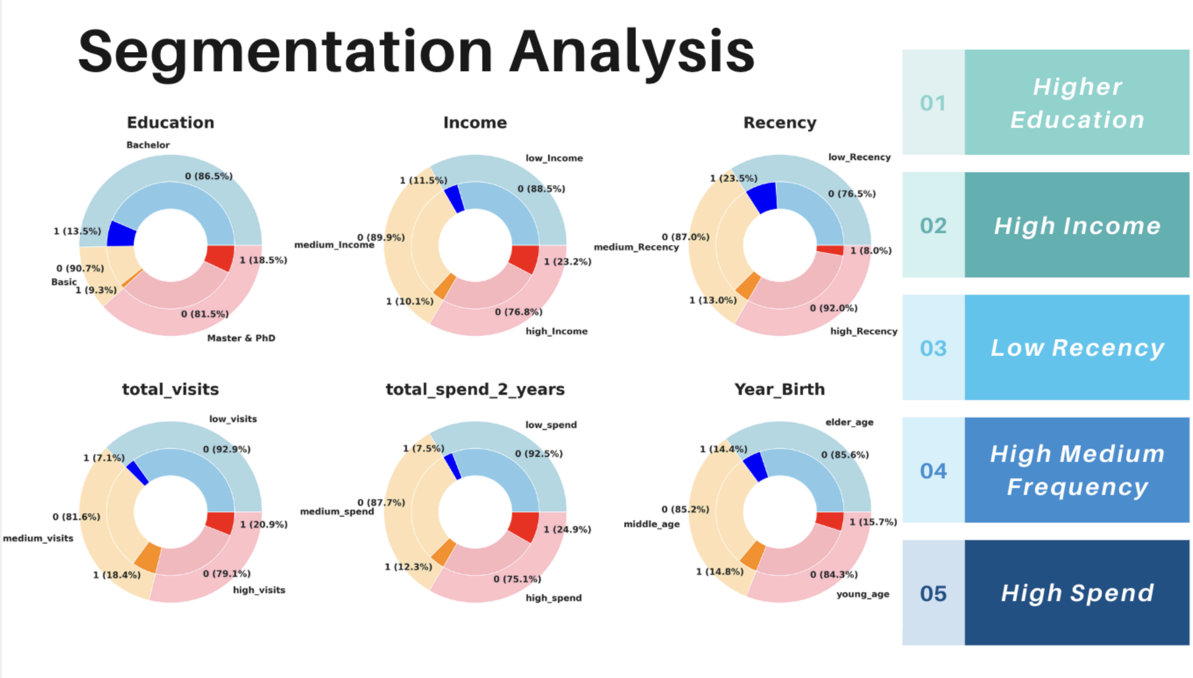
4. Results and recommendations:

We found that our model can learn from user demographic features and behavior features to predict response to the marketing campaign effectively.
Logistic regression has a low precision and high recall, which says it's good at including the people who are going to respond, but the prediction is not precise, meaning that it classified a lot of non-responders as responders, which could be a problem if we have limited marketing budget and human resources, and we would waste money by targeting people who are not going to convert.
XGBoost model has a more balanced precision and recall, the recall is not as high as logistic regression but the precision is higher than logistic regression, and has a higher AUC. Based on how much our marketing budget and goals are, we can adjust classification threshold to trade off precision and recall.

As the SHAP values indicated, features such as education level, marital status and recency were strong predictors of the XGBoost model - can be used to target potential users.

Targeting strategy:
- recommend that marketing be focused on single, higher educated customers
- engage with customers who have recently shopped at the supermarket
- personalized messages/calls
- consider interactions between variables (can be done in version 2)

We also include a plan B when we don't know the user demographics just in case, to account for the scenario where we cannot use the machine learning models we build. We conducted segmentation analysis, and used solely user behaviors data in our store to find the group of people with a high conversion rate.
We find that **people with low recency, high and medium visiting frequency, and high spend in our store are going to convert more likely to our marketing campaign**, which provides the marketing team a very useful guide to target customers more precisely and effectively.



Segmentation Analysis

01 Higher Education

02 High Income

03 Low Recency

04 High Medium Frequency

05 High Spend

5. Project plan for Version 2

5.1 Future Stages

Based on our current findings, we will improve the analysis of conversion rate for the superstore marketing campaign in the following five stages in Version 2:

The first stage is feature selection. Our current models use all features for training, but there may be some features in here that are not important and even affect the performance of the model. In Version 1 we were using SHAP values to help interpret the black box model, which can help us to understand the importance of the feature. We also plan to apply techniques such as Recursive Feature Elimination (RFE) with cross-validation to identify and eliminate these unimportant features.

The second stage is to develop advanced models and further evaluation. The logistic regression model has the problem of low precision, while XGBoost has the problem of overfitting. So developing more advanced models is a potential solution to better balance high performance and generalization. In addition, we also consider using K-fold cross validation to ensure that the model performs reliably and reproducibly on different datasets.

In the second version, we will also attempt to perform causal inference and external validation. Although our current models give us some insights like people with high education level are more likely to respond to campaign offers, we cannot infer whether this correlation is also causal. We plan to conduct external validation through A/B testing to delve deeper into the reasons behind higher conversion rates, so that we can better understand user behavior as well as provide insights for future marketing campaigns.

Next stage is continuous improvement and iteration. As user data increases and new users join, we need to ensure that the model can maintain continued good performance on new data. This stage includes creating a clear schedule for model retraining to continually tune the model.

The last stages to consider the development of dashboards. By confirming the Key Performance Indicators (KPIs) with the marketing team, we plan to use tools such as Tableau or PowerBI to develop interactive dashboards for real-time data monitoring and to show current sales data for non-technical stakeholder.

5.2 Deliverables

Stage 1:

- Feature importance report describing which features should be retained or removed and the reason.
- Dataset after feature selection

Stage 2:

- Robust model(s) and a summary of the model's evaluation.

Stage 3:

- A summary report containing causal inferences, A/B test design, execution and result analysis, and recommendations on marketing strategies.

Stage 4:

- Feasible model iteration plans based on past experience.

Stage 5:

- Documentation of KPIs from the marketing department to help our business intelligence engineer design dashboards.
- Interactive dashboards that display real time data

5.3 Resources

Human Resources:

- Data Engineer: Design and implement a data pipeline to ensure we can efficiently acquire new customer and sale data.
- Data Analyst: Process and explore data to provide insightful analysis for future marketing and modeling phases.
- Machine Learning Engineers: Perform feature engineering, develop machine learning models and perform parameter tuning and performance optimization.
- Risk Manager: Identify potential risks associated with the project and develop strategies to mitigate these risks.
- Financial Analyst: Control marketing budgets, perform customer acquisition cost analysis and evaluate campaign ROI.
- Marketing Specialist: Maintain good relationships with superstore customers and encourage customers to purchase our new Gold memberships.

Technical Resources:

- Visualization and dashboard development tools such as Tableau, Power BI
- Databases to store user data such as Oracle, AWS
- A/B testing related tools such as VWO (Visual Website Optimizer), Google Analytics

5.4 Potential Risk

Data Quality and Availability

- **Assumption:** The assumption of a well-established data infrastructure implies that the organization has a robust IT environment with capable systems for data storage, management, and retrieval. This includes up-to-date hardware, sophisticated software solutions, and effective data governance policies that ensure data is accessible and secure.
- **Impact and Mitigation:** A well-established infrastructure typically allows for smoother operations and less downtime, which is critical for real-time data processing and analytics. However, even in such environments, unexpected data quality issues like incomplete data sets, redundant entries, or outdated information can occur, potentially skewing analysis results. Our mitigation strategy, therefore, focuses on preemptive data quality checks and continuous monitoring to ensure that the data used in modeling is accurate and representative of the customer base.

Modeling Accuracy

- **Assumption:** Assuming customer appeal and pricing sensitivity refers to the belief that the underlying dynamics of customer behavior and their response to pricing strategies are well understood and stable over time.
- **Impact and Mitigation:** This assumption can be risky if customer behavior shifts due to external factors like economic changes, new market entrants, or evolving consumer preferences. To mitigate these risks, our model validation processes, including cross-validation and the use of ensemble methods, are designed to test the model's predictions against multiple scenarios and data samples, ensuring robustness and adaptability to change.

Technical Integration

- **Assumption:** Relying on a well-established data infrastructure for the integration with existing CRM systems suggests confidence in the technical capacity of current systems to support new functionalities.
- **Impact and Mitigation:** This assumption is critical because any misalignment between new models and existing systems can lead to integration failures, resulting in delays and increased costs. The mitigation plan involves detailed assessments and close collaboration with IT teams, which is essential for anticipating and solving compatibility issues. Beta testing serves as an additional safety net, allowing for the identification and correction of technical problems before full-scale deployment.

Customer Perception and Privacy

- **Assumption:** Regulatory compliance assumes that all project activities will adhere to legal standards concerning customer data privacy and marketing practices.
- **Impact and Mitigation:** This is a foundational assumption, especially in industries heavily regulated regarding data use, such as telecommunications and finance. Non-compliance could lead to significant legal penalties and damage to the company's reputation. Our mitigation strategy focuses on transparency and communication with customers about how their data is used, ensuring they understand and consent to data practices, thus maintaining trust and compliance.

Staff Expertise and Resources

- **Assumption:** This assumes that the team has sufficient knowledge and resources to undertake and complete the project successfully. Existing expertise and established training protocols are deemed adequate.
- **Impact and Mitigation:** The assumption that current staff capabilities align with project needs is crucial for maintaining timelines and quality. However, rapid technological advancements and project-specific challenges may reveal gaps in skills or resource limitations. By continuously evaluating team capabilities and addressing gaps through training or hiring, the project aims to remain agile and adequately resourced.

By understanding and addressing these assumptions, the project positions itself to better navigate potential risks, ensuring robustness in its approach to identifying price-sensitive customers.

Critical Success Factors

- **Accurate and Robust Predictive Modeling:** The success of the project heavily relies on the ability to develop a predictive model that accurately identifies potential buyers of the Gold membership. This involves selecting the right features, using appropriate modeling techniques, and continuously monitoring and refining the model based on performance.

- **Effective Integration with CRM Systems:** Seamless integration of the predictive model with existing CRM systems is crucial for the efficient execution of the campaign. This includes ensuring that the model outputs are effectively used to target potential customers through the CRM.

- **Customer Experience and Satisfaction:** Maintaining or enhancing customer satisfaction and loyalty by making offers that are perceived as relevant and beneficial. This involves understanding customer needs deeply and personalizing communications effectively.

- **Stakeholder Communication:** Regular and clear communication with all stakeholders, including project updates and changes, to ensure alignment and address concerns promptly. Effective communication fosters stakeholder buy-in and supports smoother project execution.

5.5 Communicate Project Progress to Stakeholders

Stakeholder Communication

Regular and clear communication with all project stakeholders is vital. We will ensure that project updates and changes are communicated effectively to foster stakeholder buy-in and support smoother project execution. This includes addressing concerns promptly and keeping all parties aligned with the project goals.

Meetings

We hold regular meetings or teleconferences with key stakeholders to discuss progress, challenges, and strategic decisions. These meetings will serve as a platform to present key progress indicators, recent findings, upcoming milestones, and address any immediate needs or concerns.

Dashboards

A real-time dashboard will be developed to provide stakeholders with ongoing access to key performance indicators (KPIs), model accuracy metrics, and campaign outcomes. This tool will foster transparency and allow for quick adjustments based on real-time data.

Reports

Detailed reports, including this one, are provided at each project milestone. These reports will cover achievements, variances from the plan, insights gained, and any necessary adjustments to the strategy. This documentation will serve as a comprehensive record of the project's progress and will aid in future strategy formulation.

6. Division of labor

Name	Part
Jingchen Fu	Executive Summary, Problem Framing, part of HW3
Jiuyuan Xie	Data preprocessing, Modeling, part of HW3
Minkang Li	Risk and Project communication, part of HW2
Yi Liu	Version 2 roadmap/deliverables, Resources, HW2