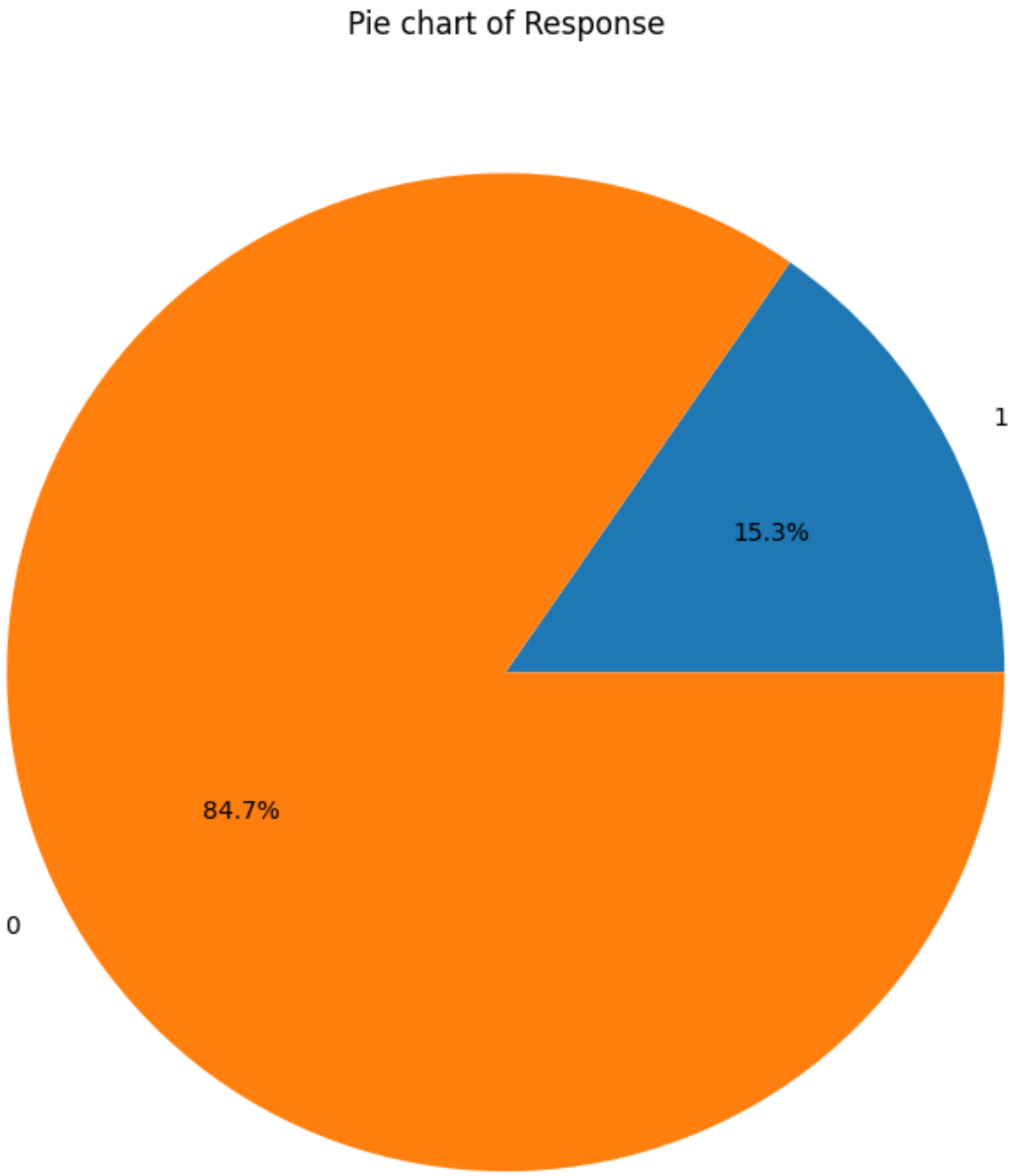
Superstore Membership Buying Decision Project

Project by Ade William Tabrani

Preface:

The author is a data scientist in a research firm, doing a custom machine learning project for a superstore client.

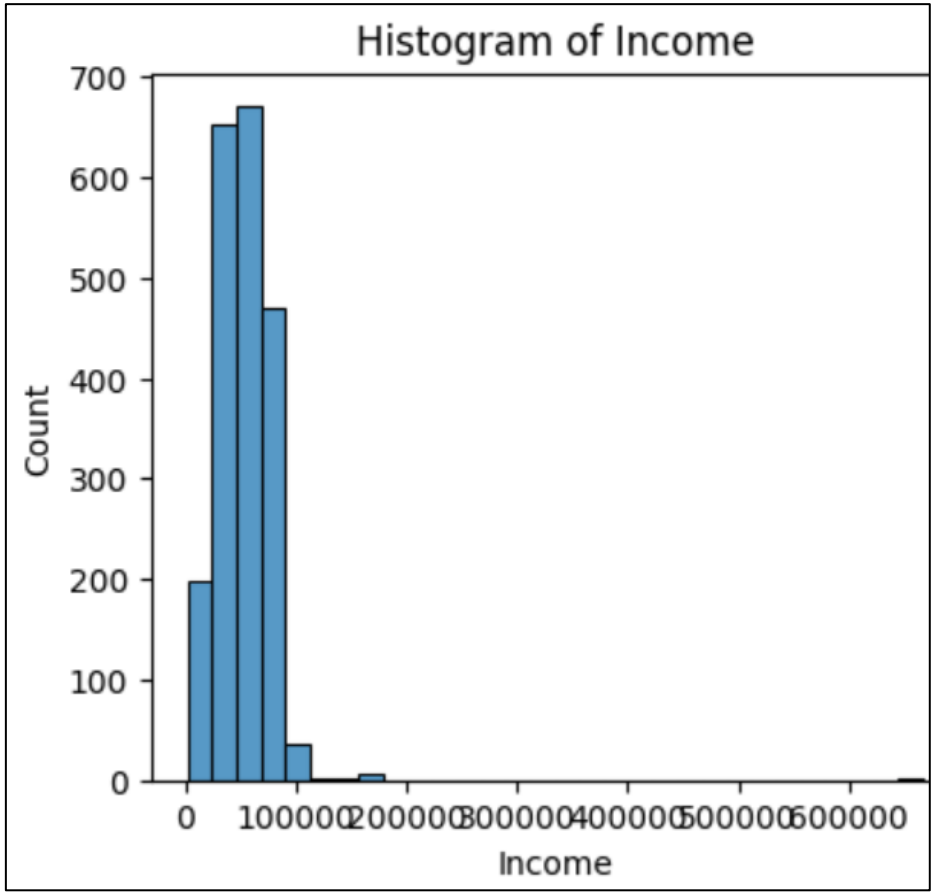
A superstore is planning for the year-end sale. They want to launch a new offer - gold membership, that gives a 20% discount on all purchases, for only $499 which is $999 on other days. It will be valid only for existing customers and the campaign through phone calls is currently being planned for them. The management feels that the best way to reduce the cost of the campaign is to make a predictive model which will classify customers who might purchase the offer.



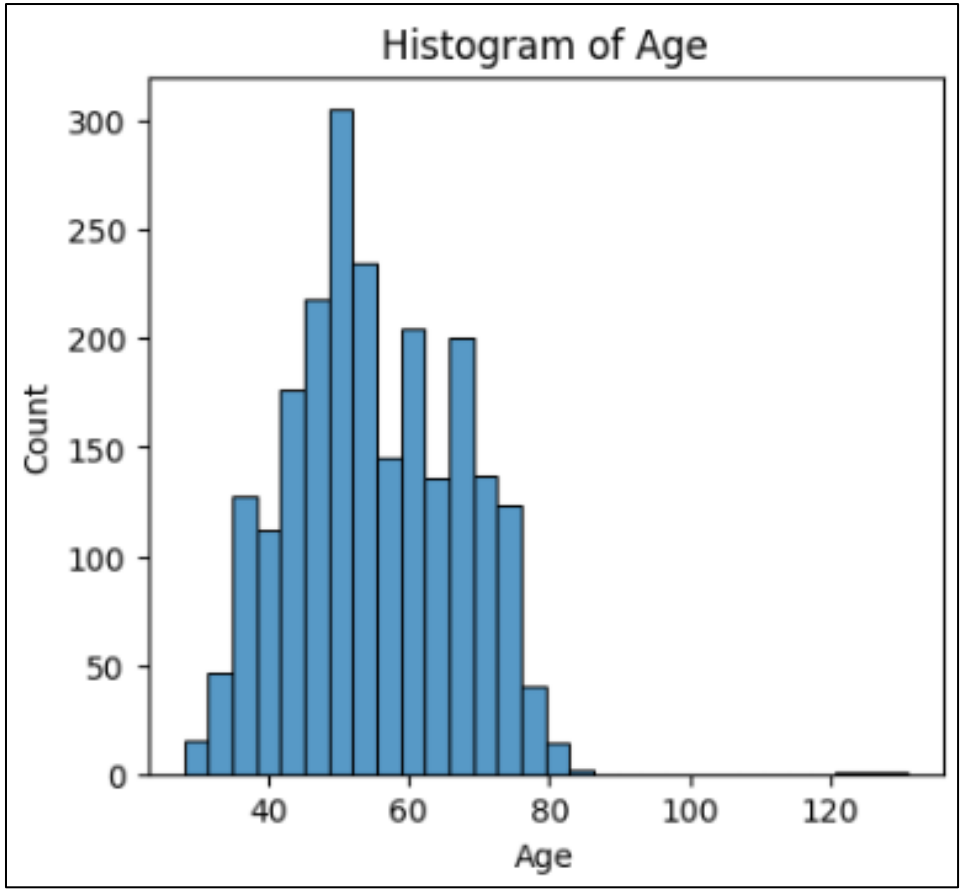
About the data:

- The data is obtained through survey at last year's campaign. It has 2,034 unique datapoints.
- Response pie chart on the left, where 1 = buy membership, 0 = does not
- Variables available in the data can be grouped into:
  - Demographical data: age, education, marital status, income, number of kids
  - Amount purchased on certain product categories
  - Number of purchases through the 3 sales channels: in-store, online, and catalogue
  - Number of complains, number of days since last purchase

Additional info on data:

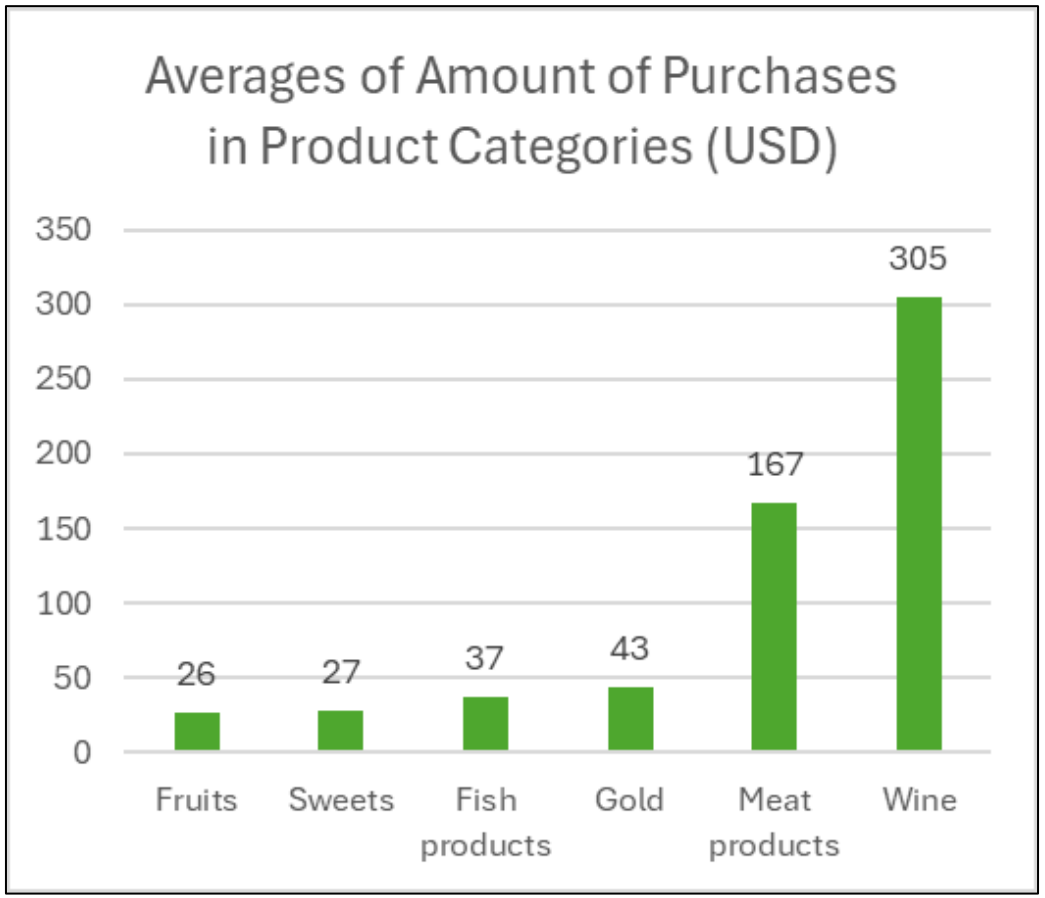


On average, customers make 52,357 USD a year. Superstore's customers are economically middle-class, where median income as of 2023 is 44,225 USD per year.

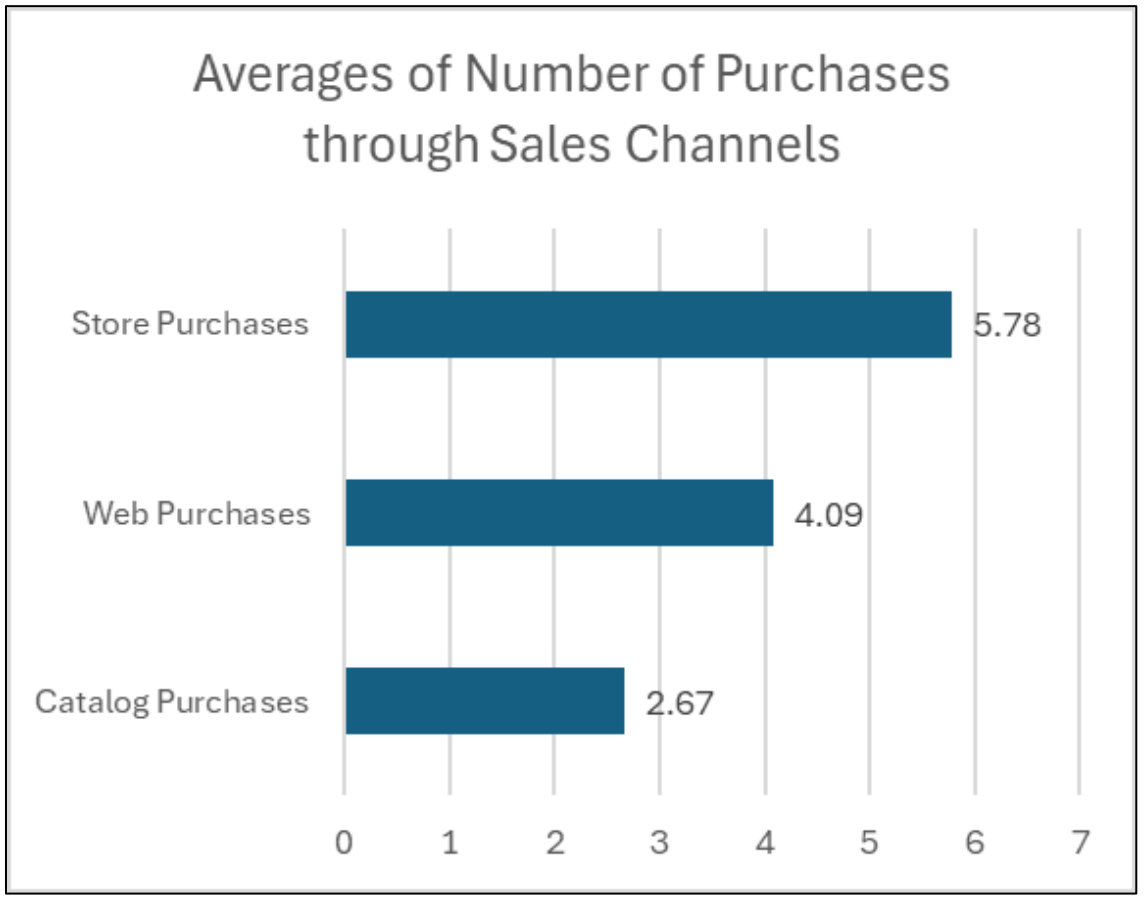


The average age of superstore's customer base is 55

Additional info on data:

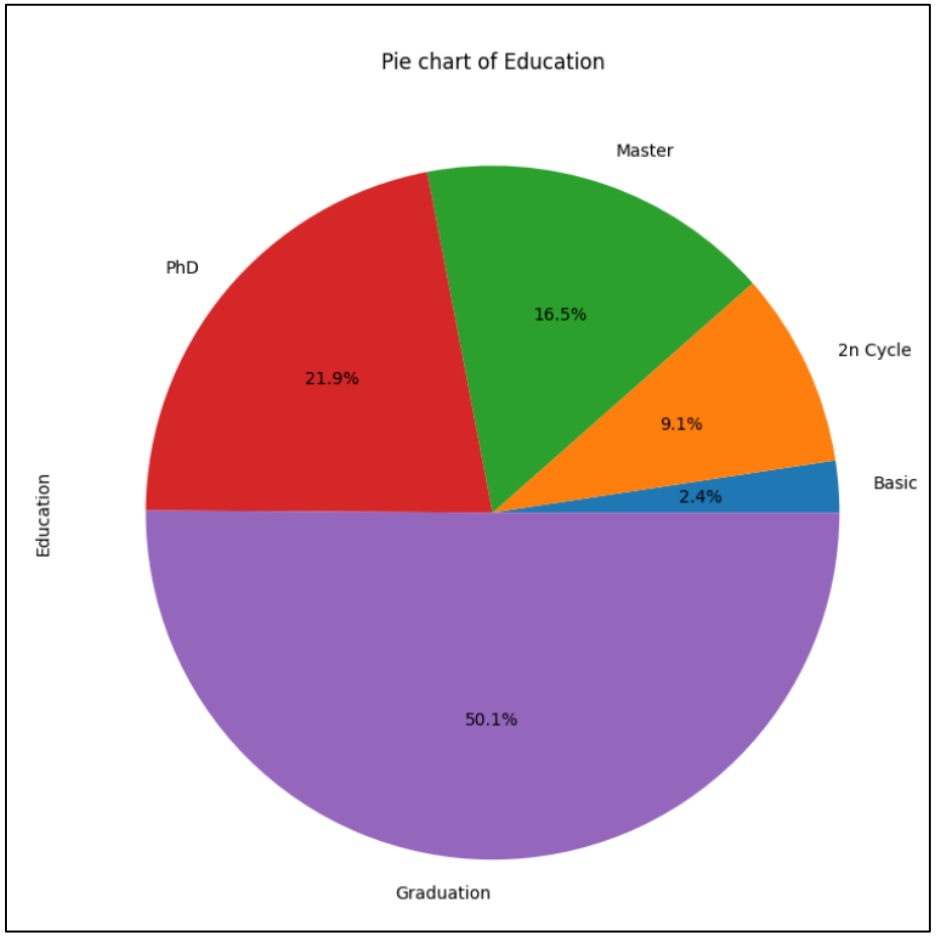


The highest value of purchases are made in wine and meat products.

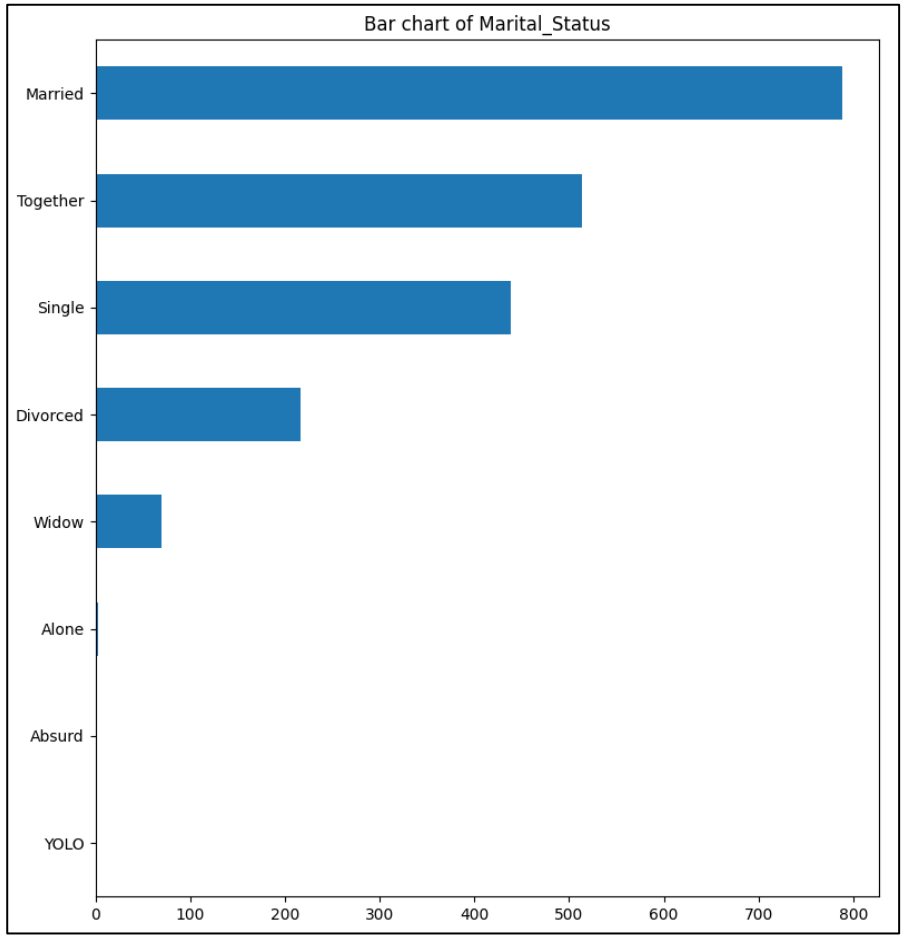


The most popular way of purchase is in-store, followed by web purchases, then catalog purchases.

Additional info on data:



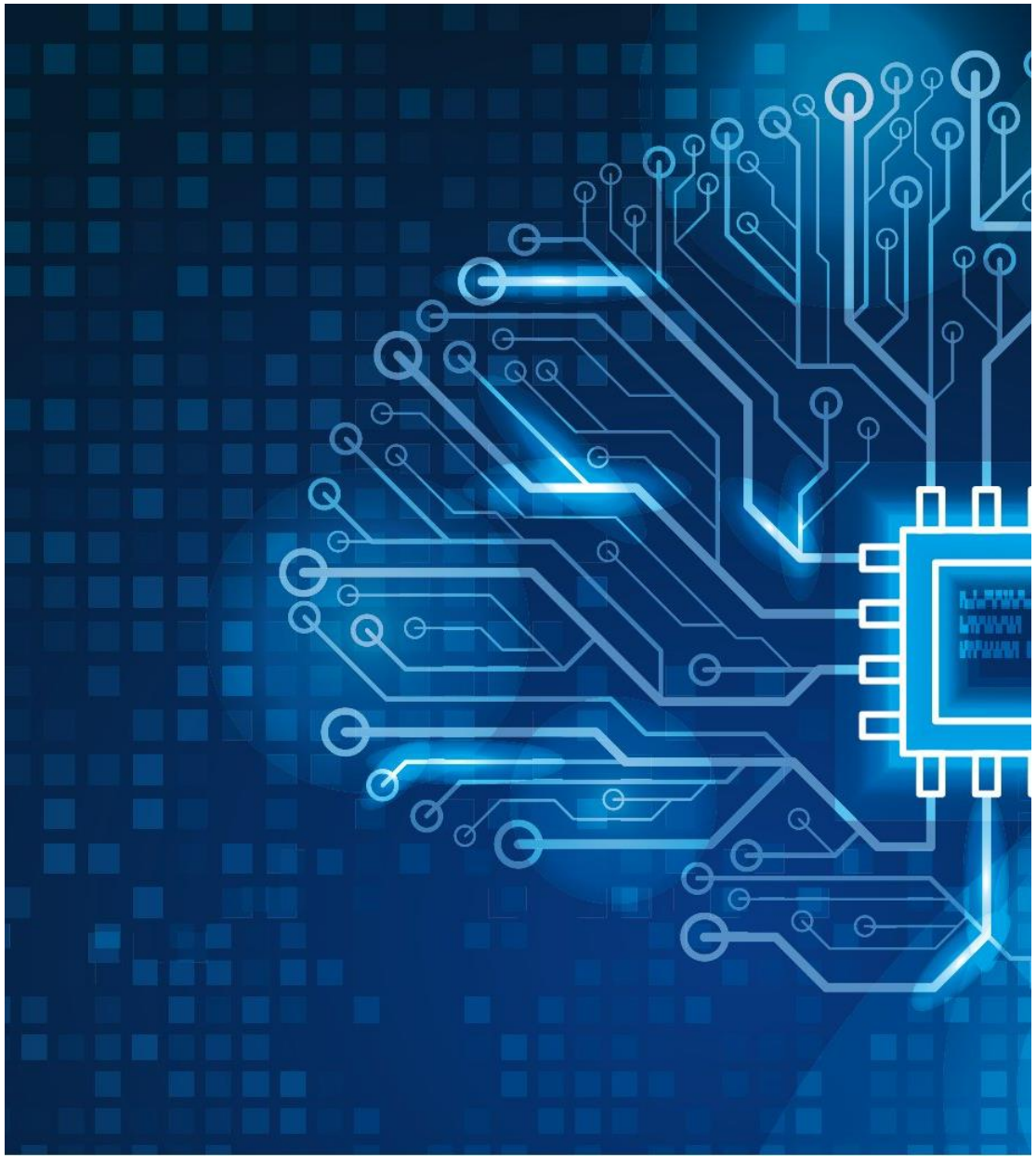
Most of the superstore customers are college graduates



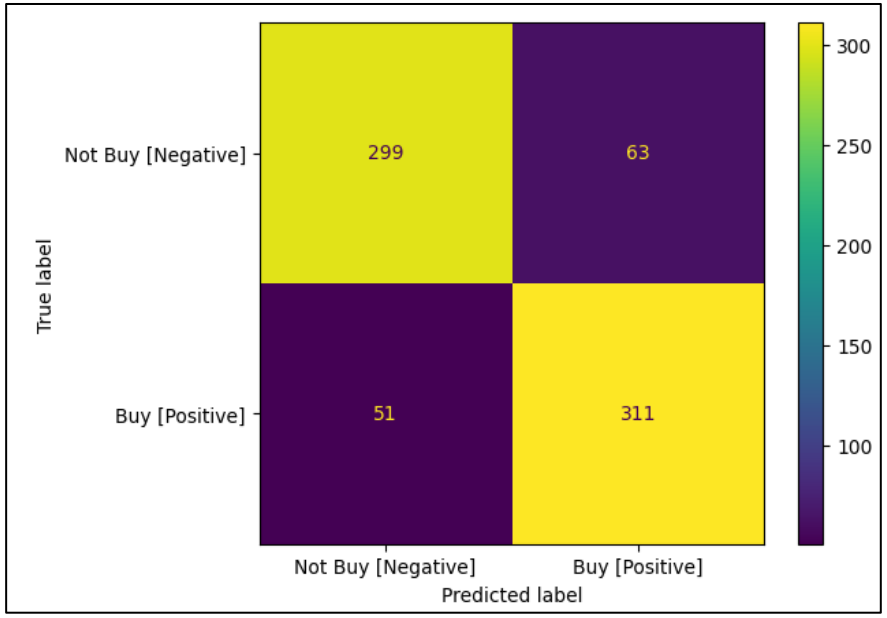
Married people made up most of superstore customers

Prediction Model

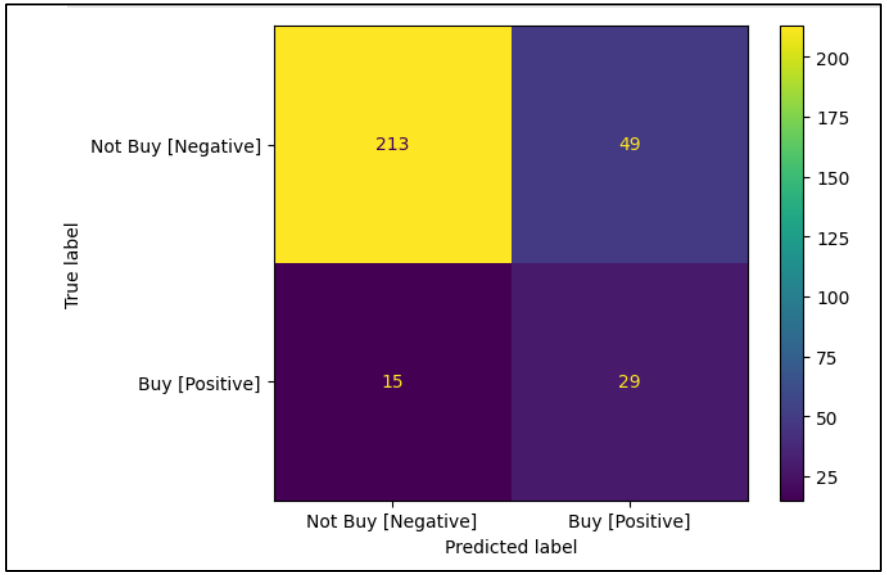
- Several models to predict between two states (in this case, buy or not buy membership) are tried, the best performing one is then picked and tuned to better its result. The model picked is named Random Forest.
- The main challenge for the machine the imbalance on number of response.



Performance of model run on balanced data:



Performance of model run on real, unbalanced data:



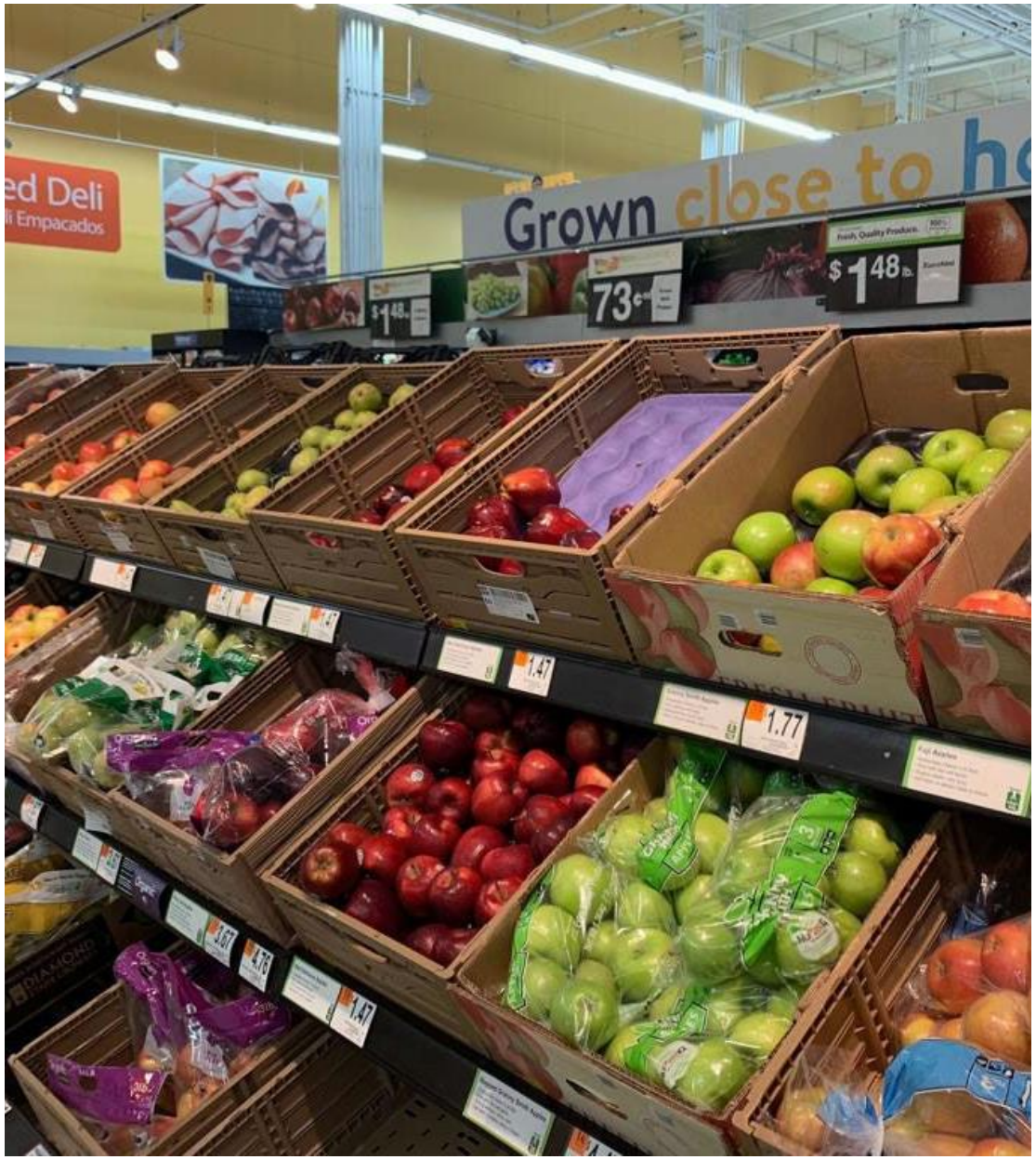
Model Evaluation:

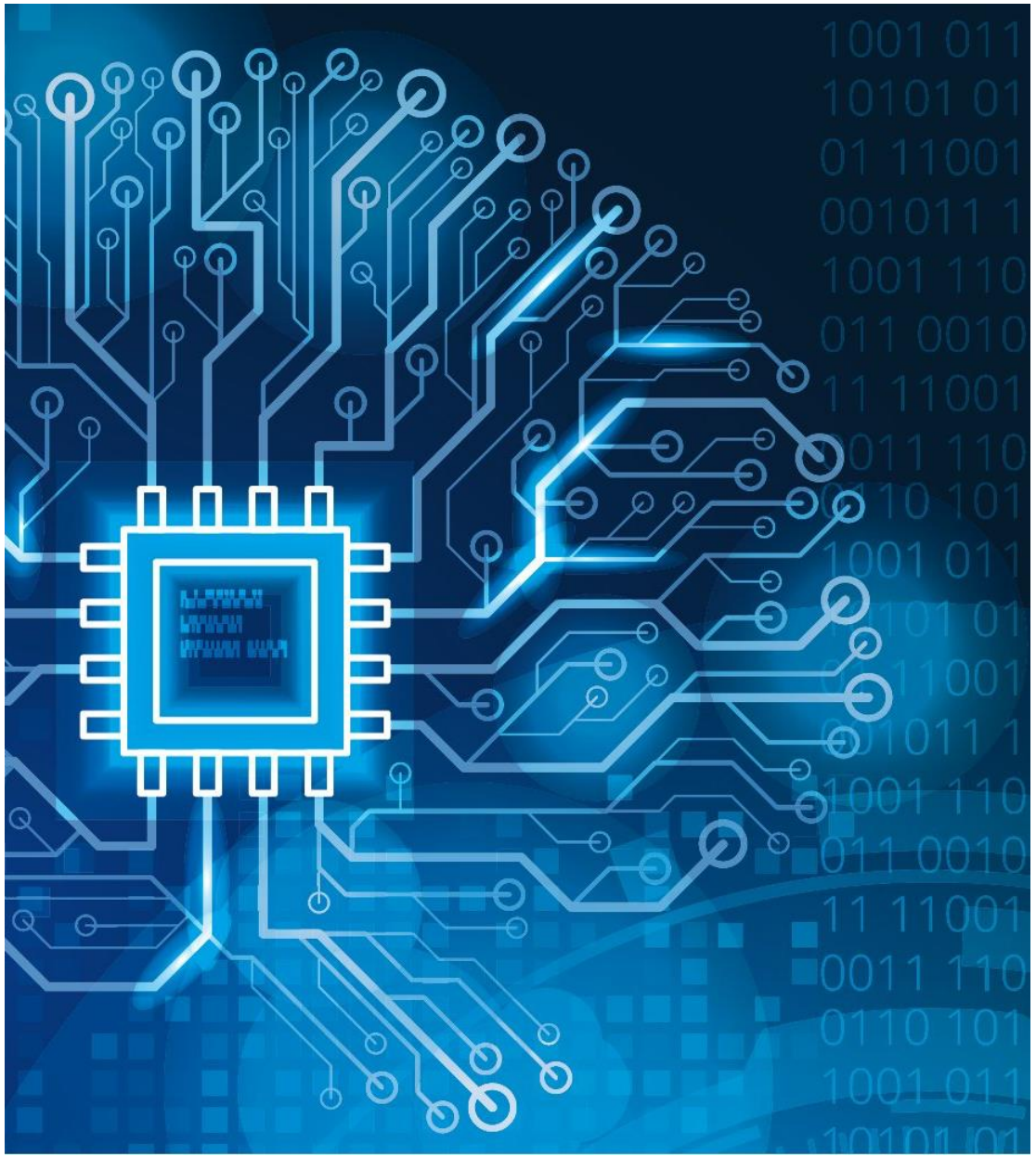
- Dataset was modified to have balanced response data, and, upon running the model with balanced data, the model was able to correctly identify/recall 86% of all membership buyers and 83% of all membership non-buyers.

- Though, when the model tries to predict using real, unbalanced data, the model was able to correctly identify 66% of all membership buyers. The model does better at predicting non-buy decision, with 81% of all non-buyer predicted correctly.

Conclusion

- Machine learning model is able to correctly identify 66% of membership buyers, decreasing the time needed for superstore’s employees to parse through each customers’ data and guesstimate which is likely to buy the membership. With this model, the marketing effort can be more easily targeted and focused to customers that are predicted to buy, decreasing marketing cost and fasten up the marketing process.
- Recommendation: Once the customers are classified into people who are likely to buy the membership, it's worth noting that building further relationship with that customer will be important to ensure their loyalty further and increase the chances of them buying the membership. This can be done through special treatment with special coupons, increased in-store interaction with, perhaps with a Customer Relationship Manager or just in-store employees in general--make them feel that they're welcomed, appreciated, and wanted. We can also give gifts, and do other customer relationship techniques.





Technical Recommendation

- Recall and F1-score can still be improved further if not for time limitations.
- There are still overfitting issues that has not been addressed
- There are still further/deeper feature selection method to do, e.g. testing out whether each feature have statistically significant relationship with the target feature
- Grid Search CV was cut down to a small number of params to save time. Further Grid Search CV can still be done to find the *best* best parameters, not to mention not all parameters are adjusted in Random Search CV and Grid Search CV.
- Grid Search CV was also not done to other models, only on final one (Random Forest), perhaps other model can do better than Random Forest upon hyperparameter tuning.
- We have not yet tested if capping the outlier will lead to a better model.
- PCA (Principal Component Analysis) was not done due to concern of PCA simplifying the information too much, we have not yet tried and compared the models if PCA was done beforehand.

For more information, contact:

Ade William Tabrani

adewilliam.tabrani@gmail.com
LinkedIn