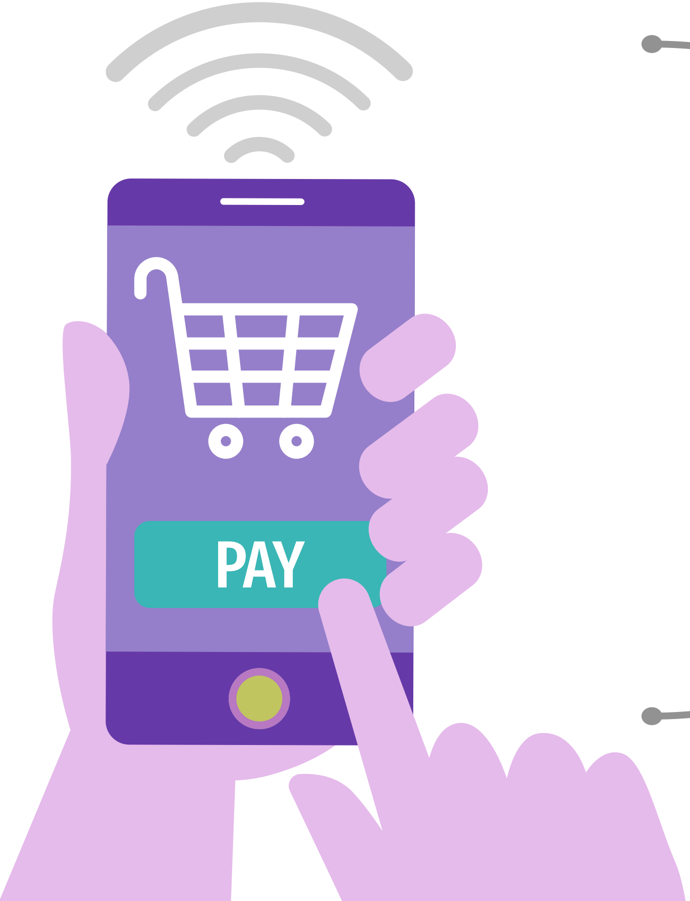
Customer Churn Prediction

Benjamin DSIF-2

Problem Statement

To prevent company from losing customer to their competitors and to retain valuable customers

Roadmap



01 Background

02 Data Exploration

03 Data Modeling

04 Conclusion

Background

- ❑ Singapore is one of the biggest e-commerce markets in Southeast Asia despite being a small island nation.

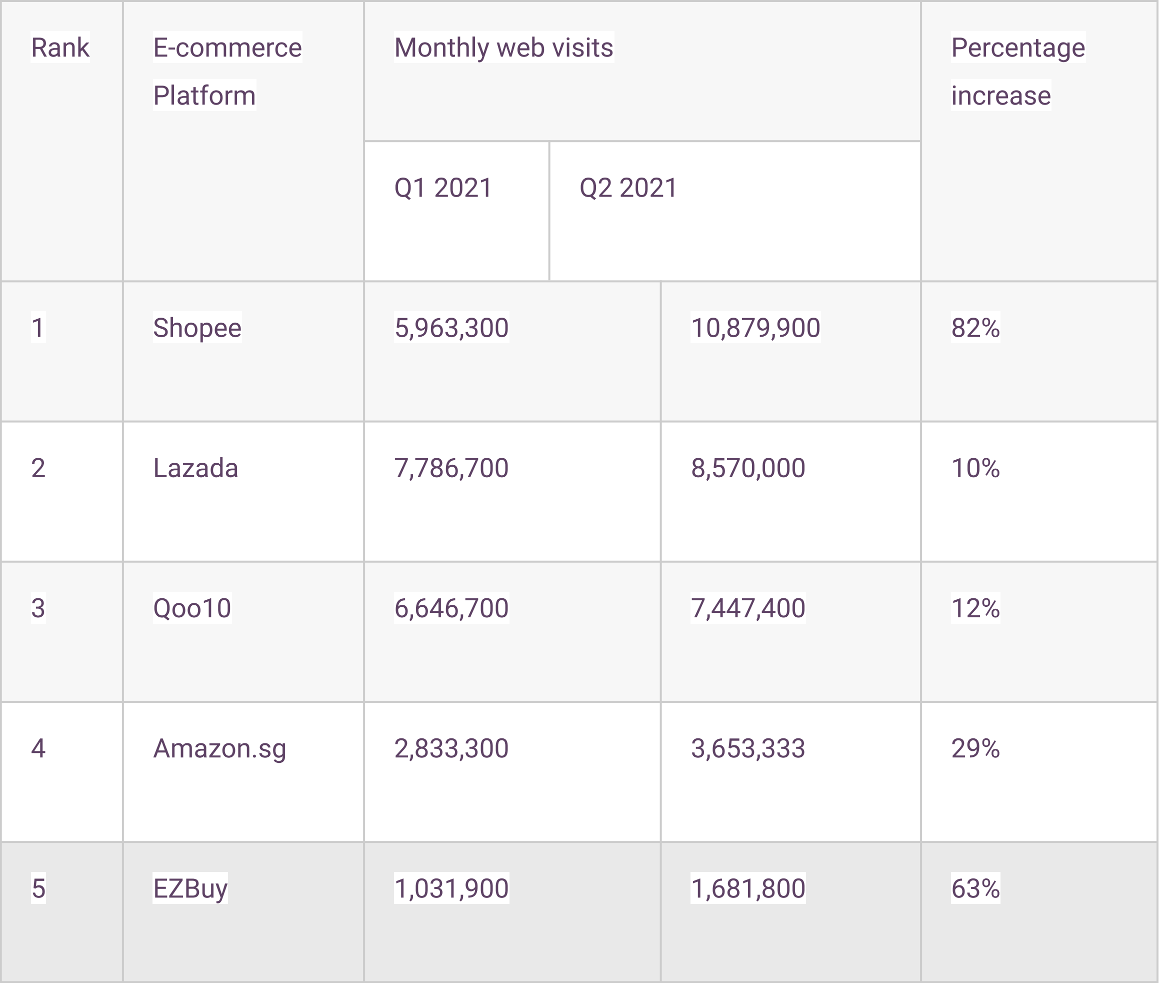
- ❑ E-commerce sales in Singapore are expected to grow to US$10 billion (S$13.4 billion) by the end of 2026

- ❑ South-east Asia's e-commerce gross merchandise value is expected to nearly double to US$254 billion in five years, up from the US$132 billion forecast in 2021, said the latest Sync South-east Asia Report

- ❑ Online sales in Singapore will hit an estimated US$8 billion this year.

https://www.straitstimes.com/business/economy/spore-e-commerce-sales-forecast-to-hit-134-billion-by-2026-report

Background



Rank	E-commerce Platform	Monthly web visits		Percentage increase
		Q1 2021	Q2 2021	
1	Shopee	5,963,300	10,879,900	82%
2	Lazada	7,786,700	8,570,000	10%
3	Qoo10	6,646,700	7,447,400	12%
4	Amazon.sg	2,833,300	3,653,333	29%
5	EZBuy	1,031,900	1,681,800	63%

https://heysara.sg/statistics-on-e-commerce-landscape-in-singapore/

Background (Customer Churning)

Why Is It Necessary?

Having the ability to accurately predict future churn rates is necessary because it helps your business gain a better understanding of future expected revenue.

In addition, when you're able to use churn prediction to forecast the potential churn rate of a particular customer, it allows you to target that individual in an attempt to prevent them from discontinuing their subscription with you.

And, since the cost of acquiring a new customer is 5x higher than keeping an existing one, there's plenty of revenue-based reason to do everything in your power to keep those existing customers.

CUSTOMER ACQUISITION
VS.
RETENTION COSTS
It costs five times as much to attract a new customer, than to keep an existing one
$
$$$$$

https://baremetrics.com/academy/churn-prediction-can-improve-business#:~:text=Having%20the%20ability%20to%20accurately,understanding%20of%20future%20expected%20revenue.&text=Predicting%20churn%20rates%20can%20also,where%20customer%20service%20is%20lacking.

Background

This Dataset (2015 - 2017) belongs to a Machine Learning Challenge hosted at HackerEarth.

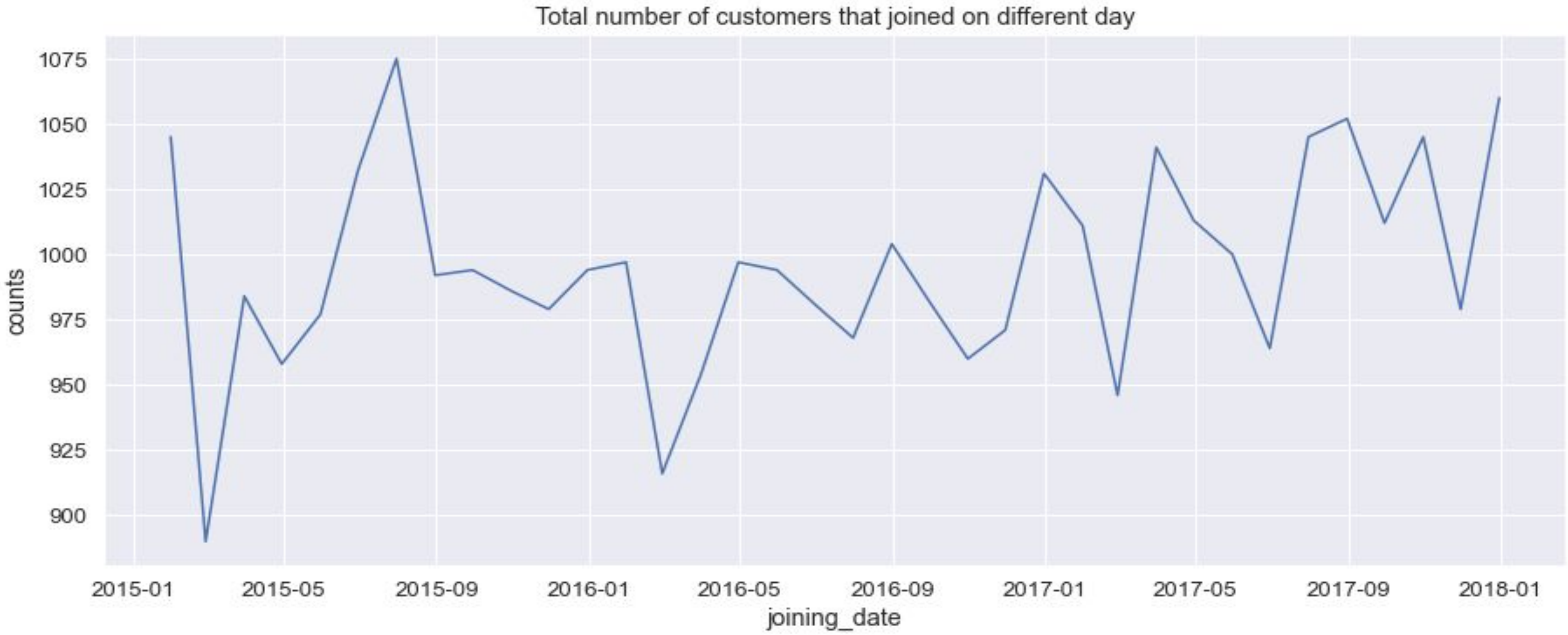
Objective : Predict the churn risk rate

Churn rate is a marketing metric that describes the number of customers who leave a business over a specific time period. Every user is assigned a prediction value that estimates their state of churn at any given time. This value is based on:

- User demographic information
- Browsing behavior
- Historical purchase data and other information

The values assigned are between 1 and 5, with 1 being the least probability and 5 being the highest probability

Data Exploration (Trend of customer base from 2015 - 2017)

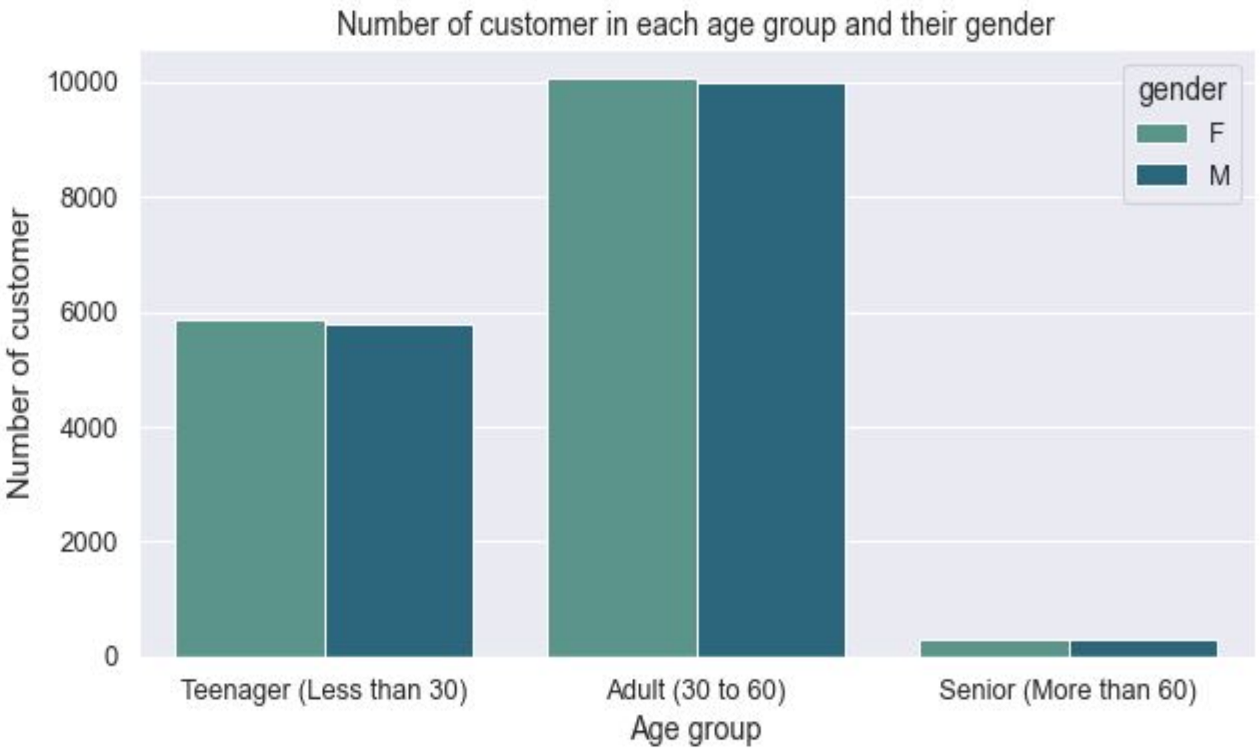


From 2015 to 2017, the number of customers who joined are fluctuating, with the highest count of 1075. However, from 2016 to 2017, a gradual uptrend can be seen which shows the number of customers are increasing.

Data Exploration (User Demographic)

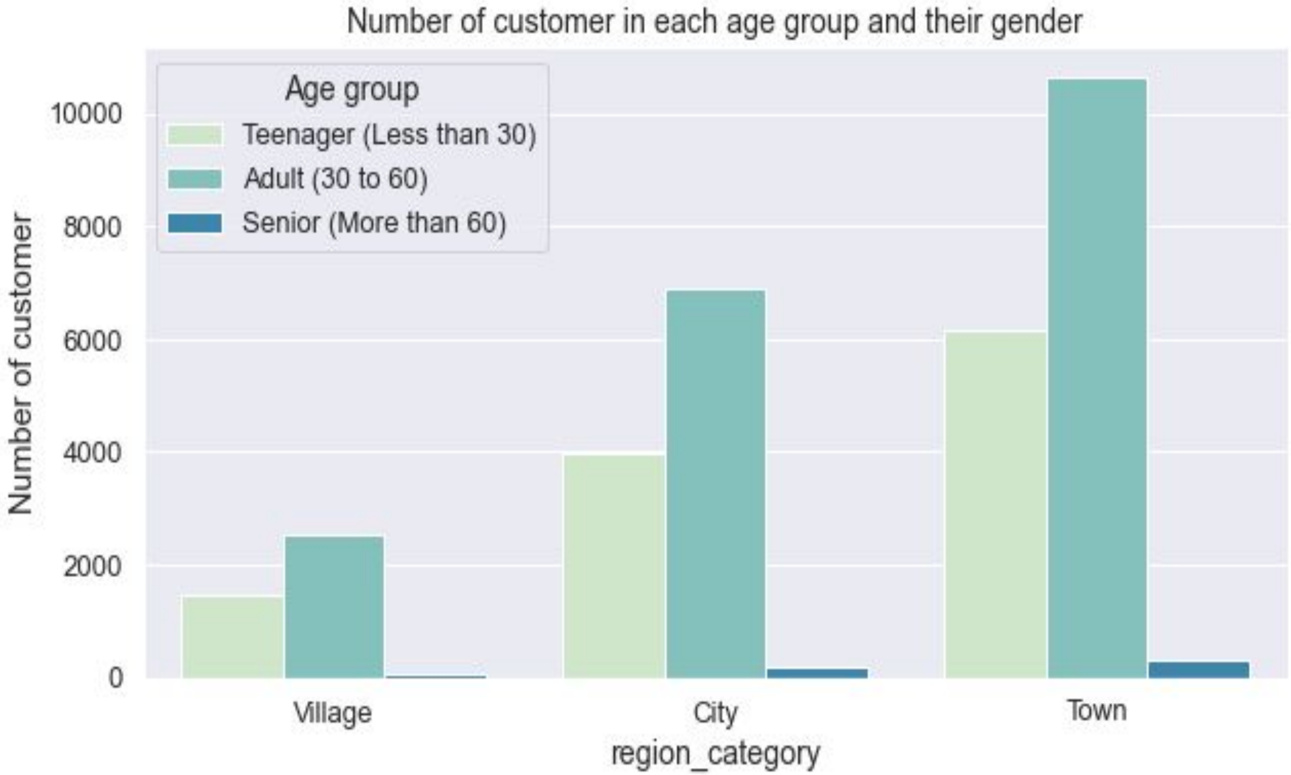
- ❑ Majority of the customers are Adult, follow by Teenager and a very small portion of Senior

- ❑ The proportion of gender are equally distributed across all age group

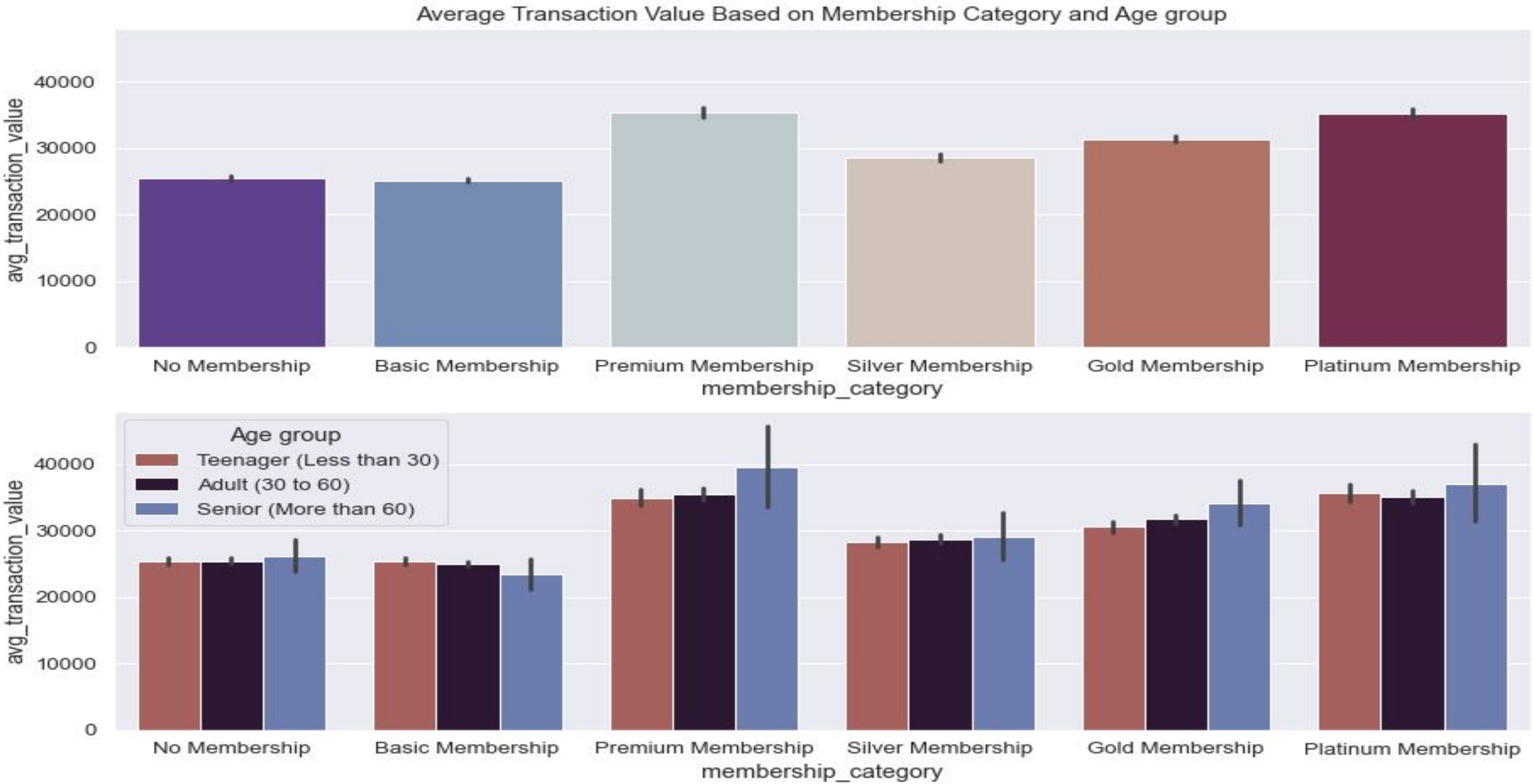


Data Exploration (User Demographic)

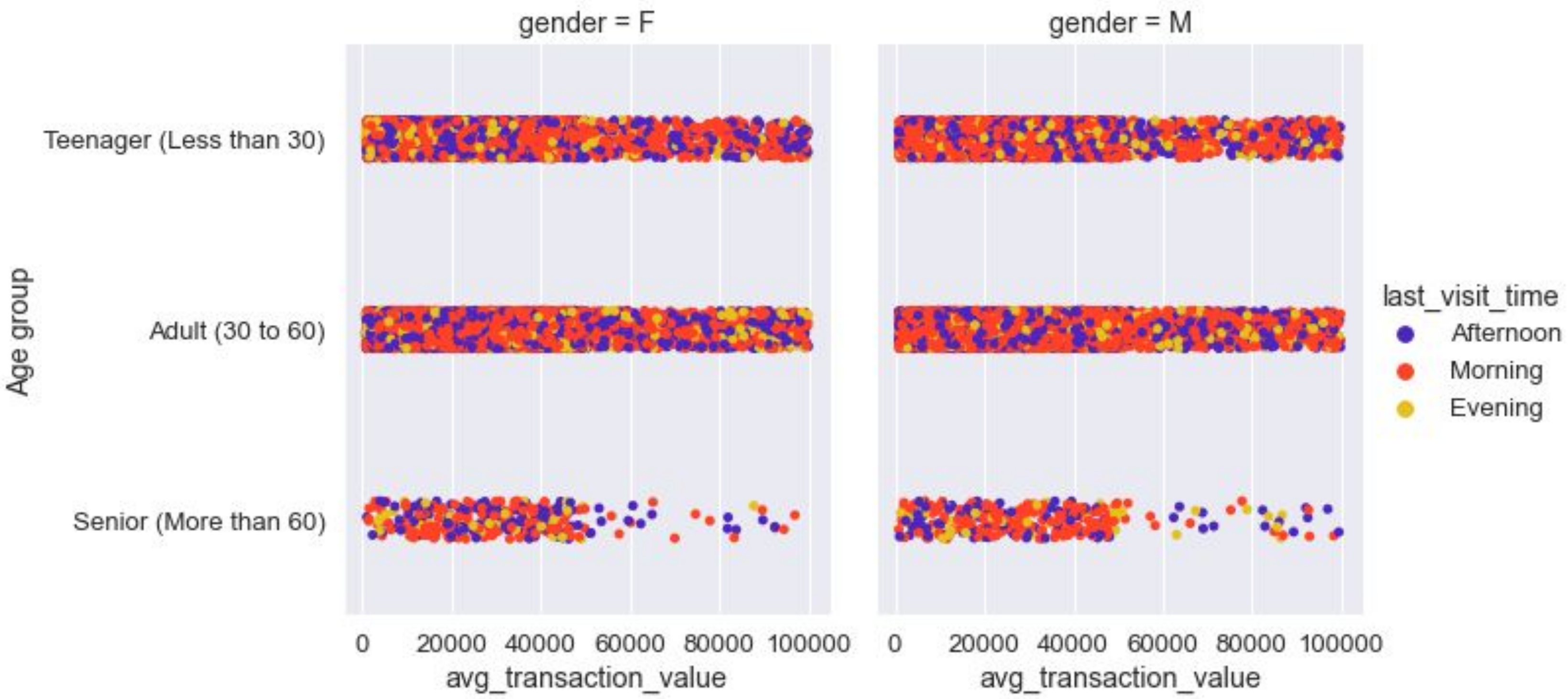
- ❑ Out of the age group, majority of the customer stays in City or Town.
- ❑ The proportion of age group are quite similar, where close to 50% of the total number of adults are teenagers



Data Exploration (User Demographic)

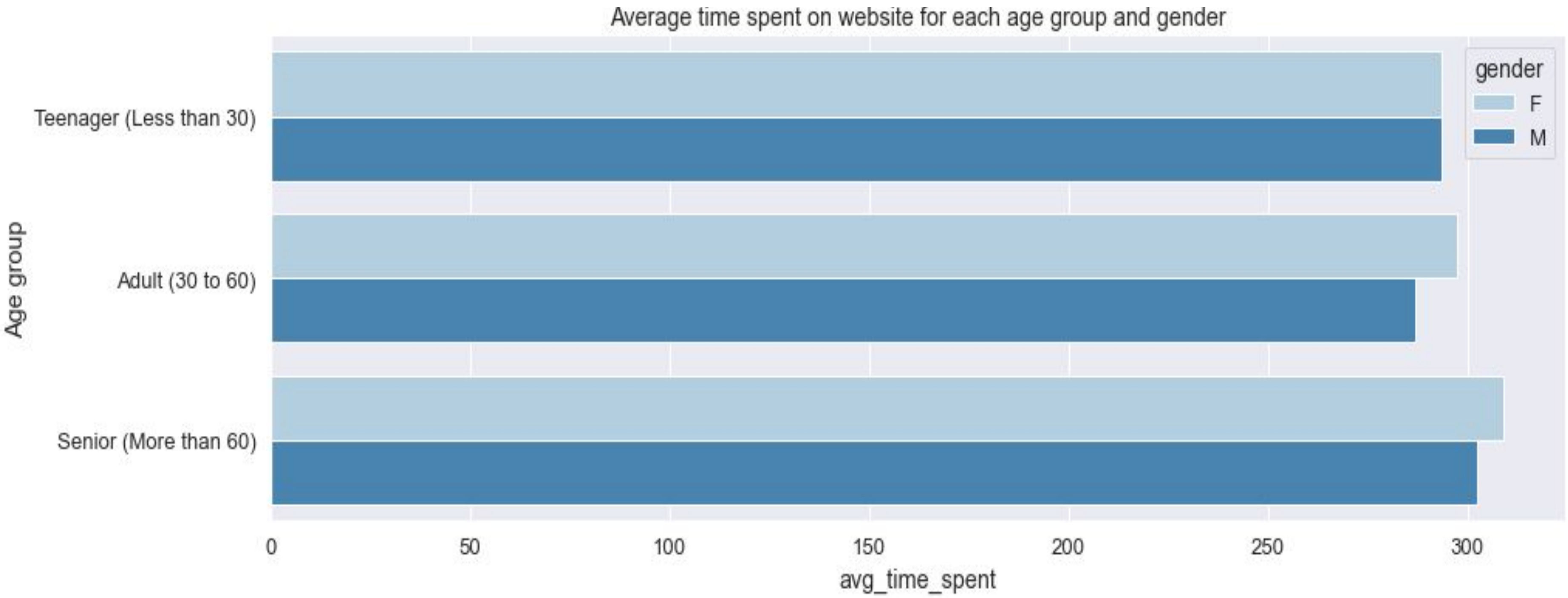


Data Exploration (Browsing Behavior)



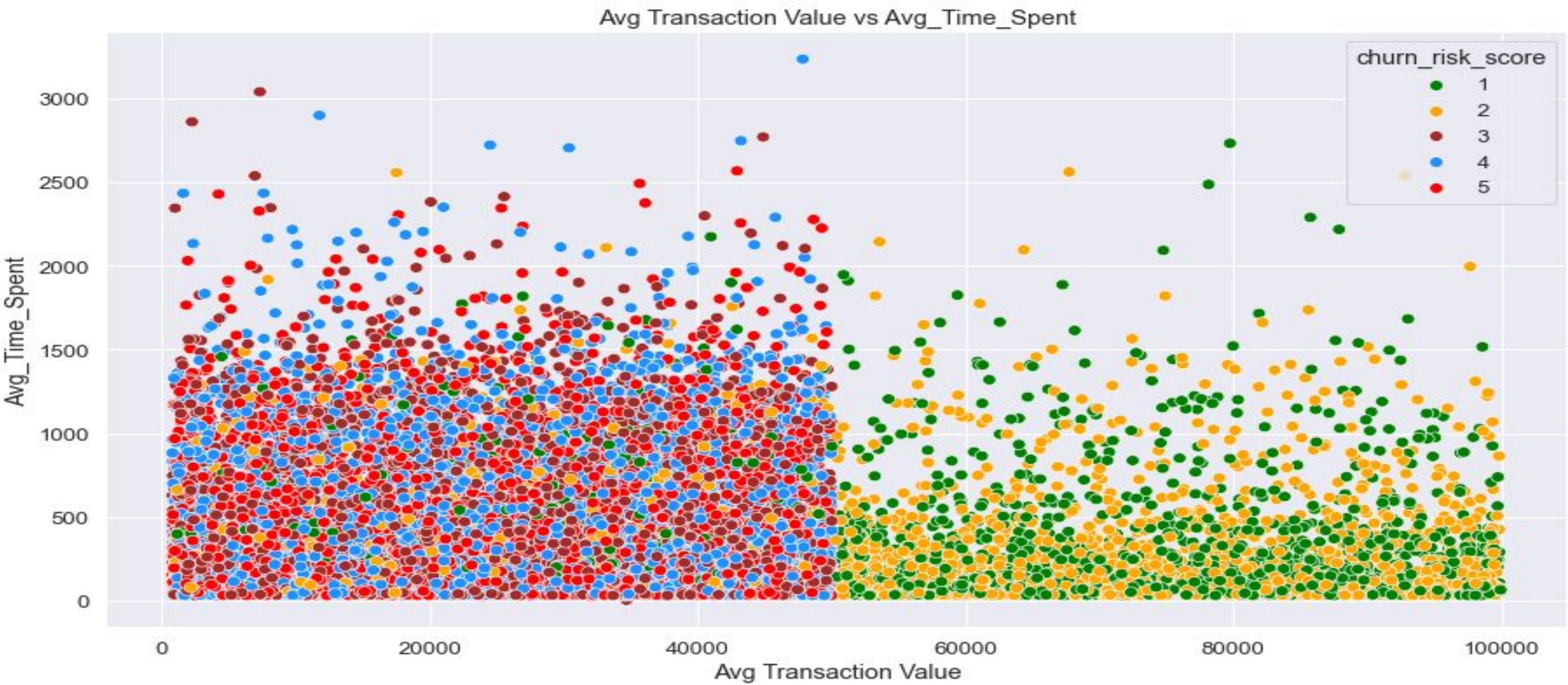
Most of the customer across all age group and gender, visits the ecommerce shop in the morning. Male senior tend to be the bigger spender as compared to female and they visit more often in the evening.

Data Exploration (Browsing Behavior)



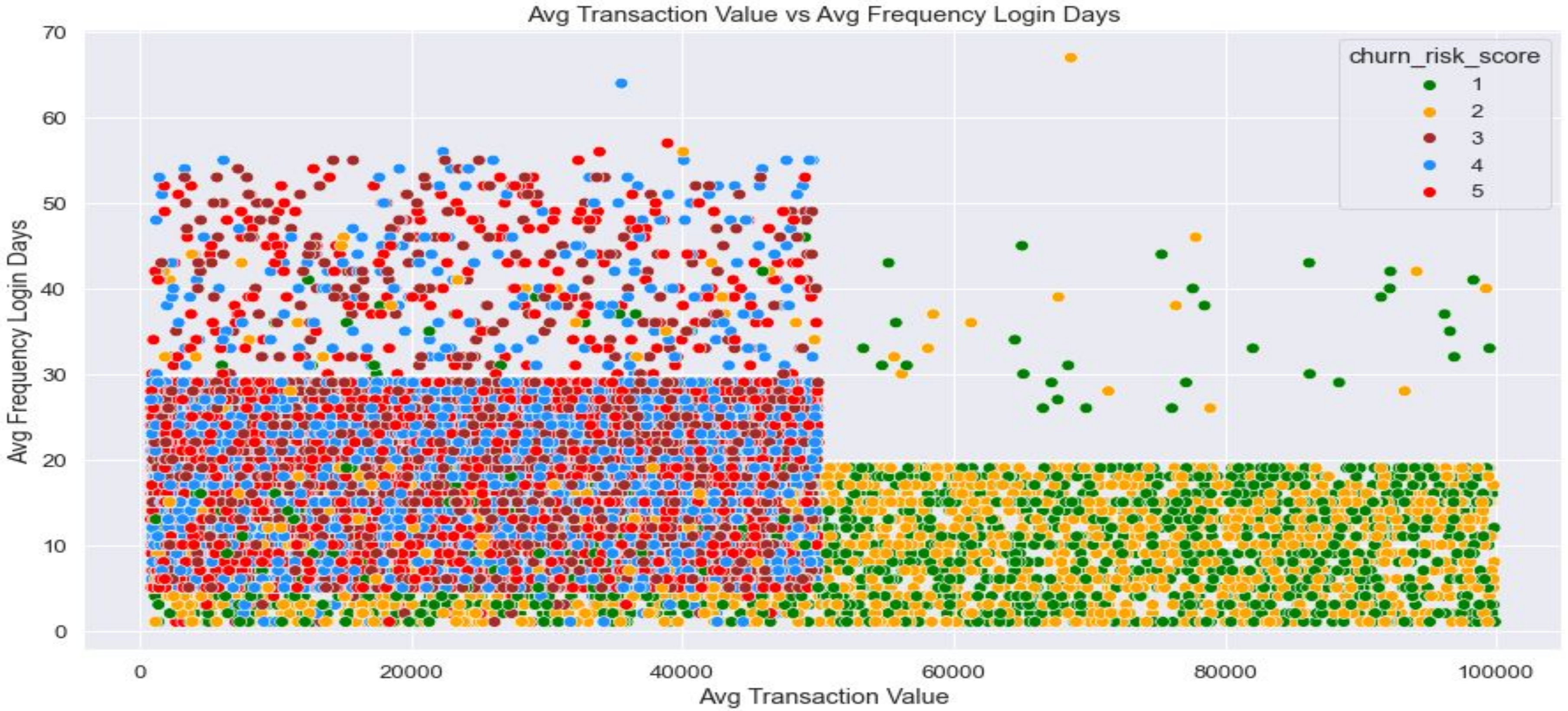
Seniors and female tend to spend more time on the website

Further Data Exploration (Avg_Transaction_Value vs Avg_Time_Spent)



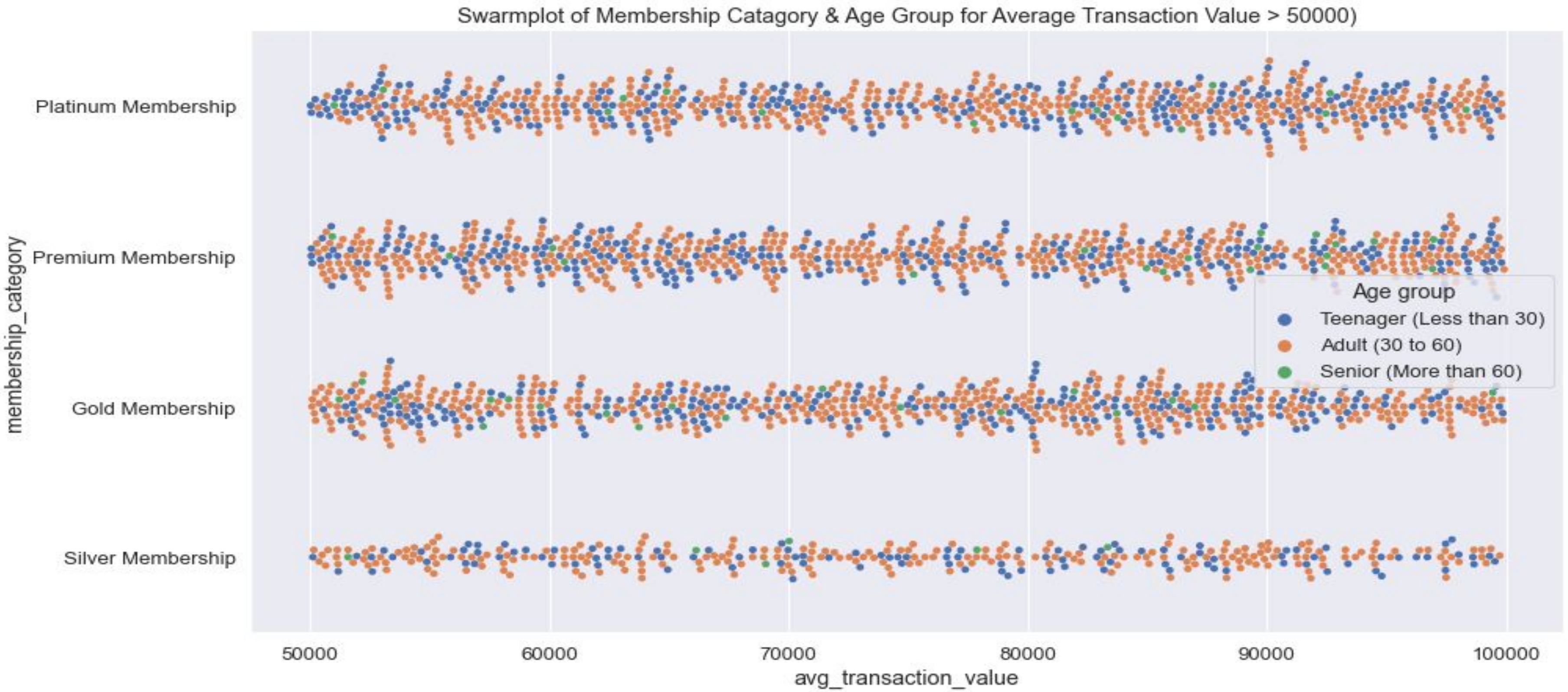
For customer whose average transaction value is more than 50000, their churn risk score is the lowest. Customer who spend more time on the website does not mean they spend more.

Further Data Exploration (Avg_Transaction_Value vs Avg_Frequency_Login_Days)



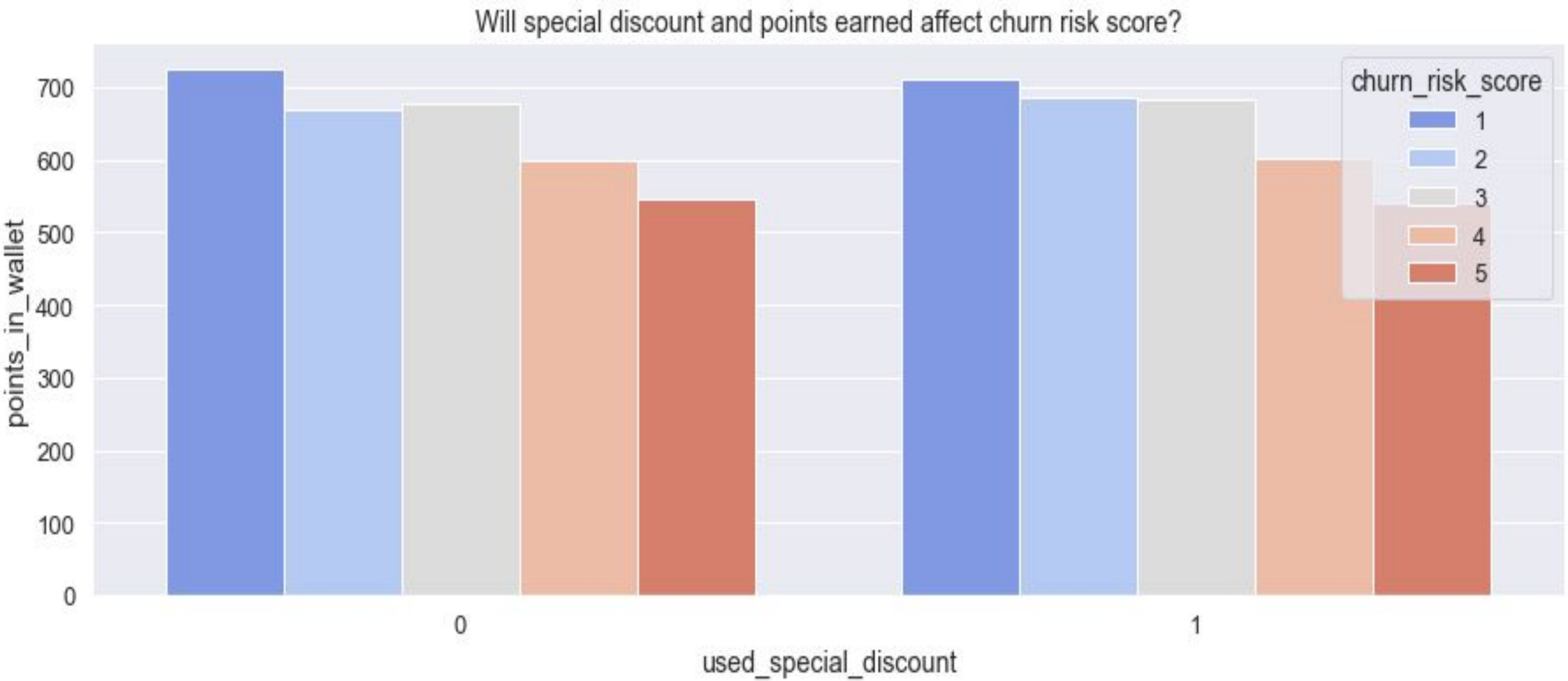
On average, customer login about 20 times to the website whose average transaction value is more than 50000, their churn risk score is the lowest.

Further Data Exploration (avg_transaction_value > 50000)



Customer who spend more than 50000 in values are from Premium, Silver, Gold and Platinum membership and majority are adults, follow by teenager with very few are seniors.

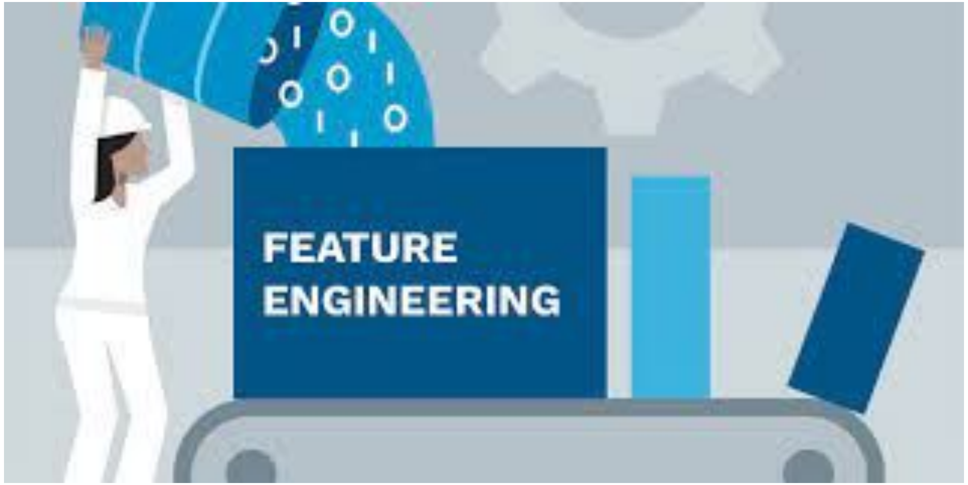
Further Data Exploration (Special discount and point earned)



Special discount does not really affect the churn risk but for ‘points in wallet’, the highest churn risk score tend to have lesser points in their wallet.

Feature Engineering

- Created Age group', eg. Teenager, Adult, Senior from ‘age’
- Created 'Time of the day', eg. Day, Afternoon, Evening from 'last_visit_time'
- Convert 'feedback' to whether if it is a positive or negative feedback
- Created 'Number of months as member' from 'joining_date'



Data Modeling Results

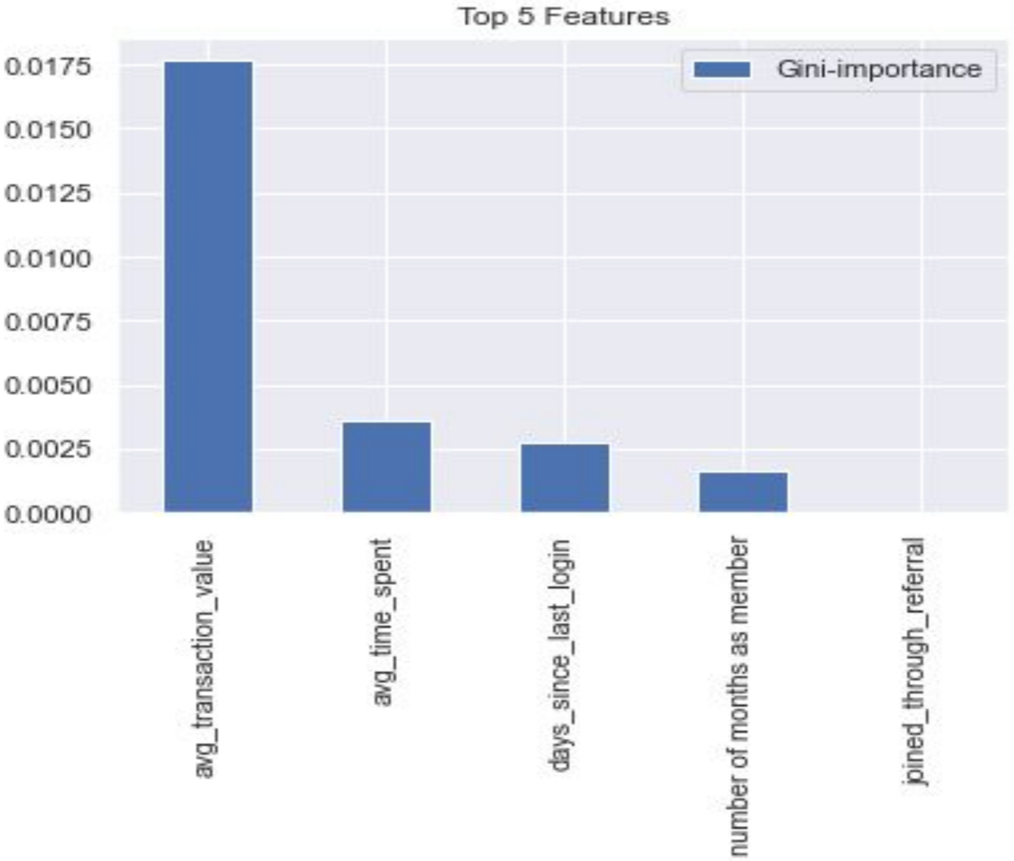
	Accuracy	Recall	Precision	F1 Score
RandomForestClassifier	0.724	0.724	0.762	0.708
DecisionTreeClassifier	0.738	0.738	0.759	0.735
Gradient Boosting	0.741	0.741	0.749	0.724

DecisionTreeClassifier is selected as the best model for the best F1 Score

Feature Importance

The top 5 features selection are:

1. avg_transaction_value
2. avg_time_spent
3. days_since_last_login
4. joined_through_referral
5. number of years as member



Conclusion and Future works

Conclusion

- ❑ The amount that customer spend on the website is the top feature for churn prediction
- ❑ DecisionTreeClassifier is the best model and not overfitted

Future works

- ❑ Use other algorithms like XGBoost, etc.
- ❑ Hyperparameter tuning, gathering data and more feature engineering to improve model score
- ❑ Deploy model for usage

Thank you