STAT 244-SC Final Project

Robin Tran

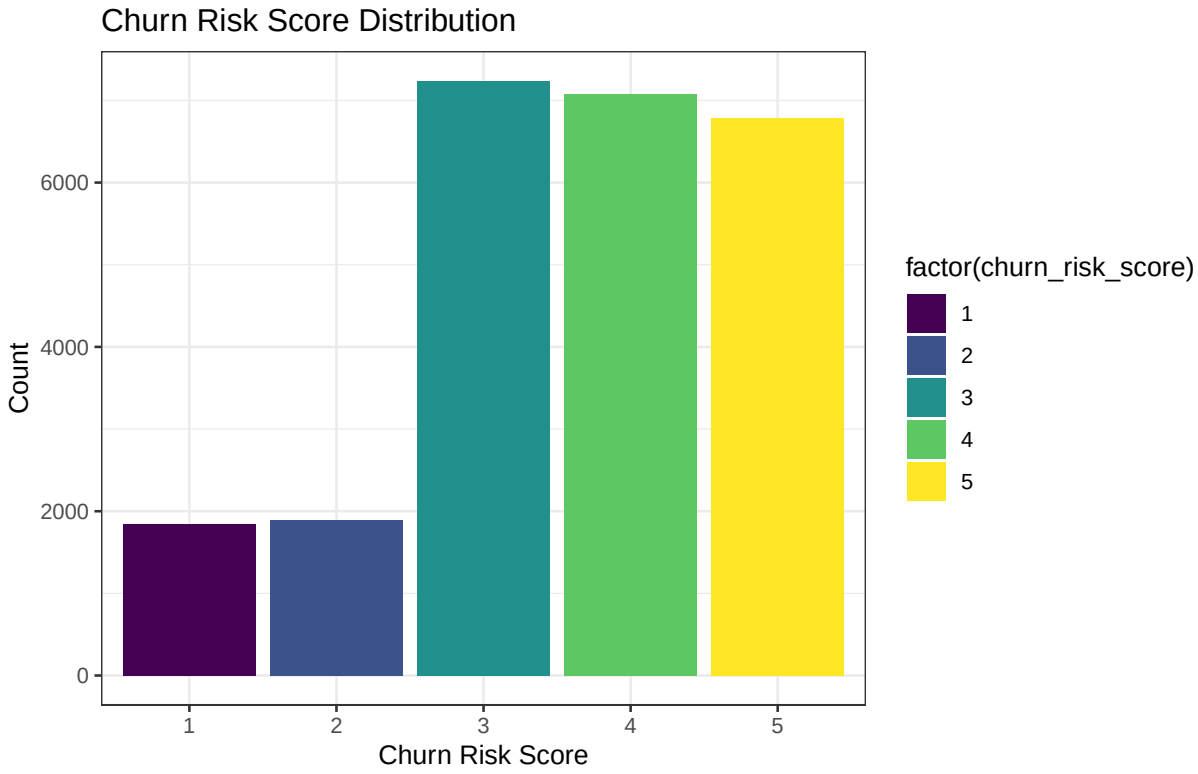
Introduction

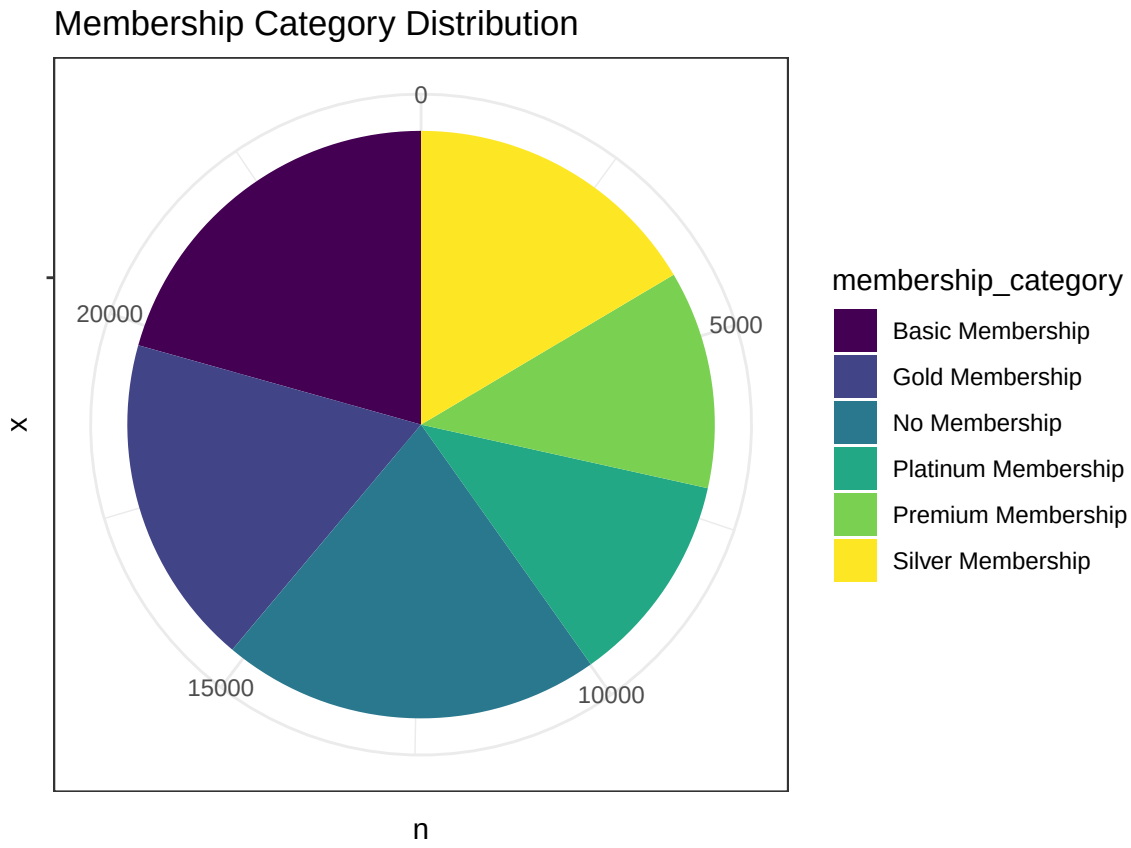
This project explores customer behavior using a dataset with demographic, transactional, and engagement features. There are two main sections. In the first section, we implemented linear regression models to predict each customer's average transaction value, comparing different model specifications using 10-fold cross-validation. In the second section, we used logistic regression to classify whether a customer is at high churn risk based on selected predictors. The goal is to understand what factors are associated with customer spending and retention, and to evaluate model performance using appropriate validation techniques.

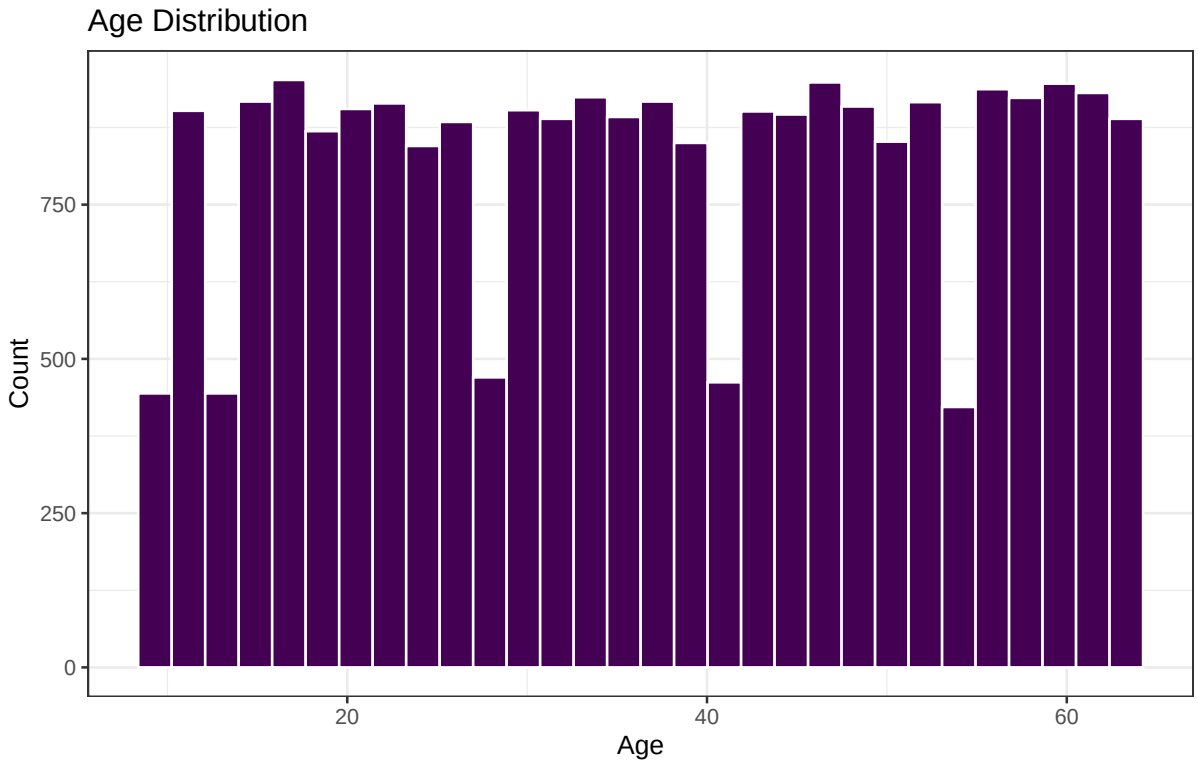
Each row in the data set represents one individual customer who has engaged with the platform and has at least one recorded purchase.

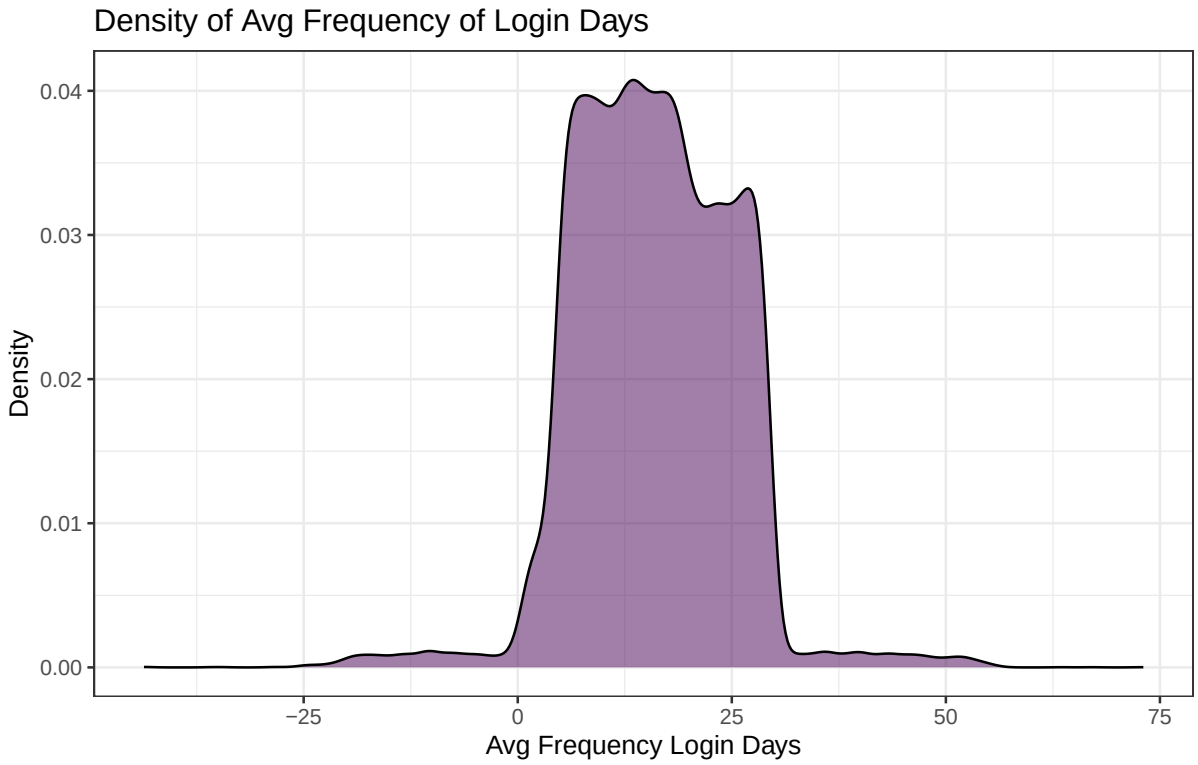
Exploratory Data Analysis

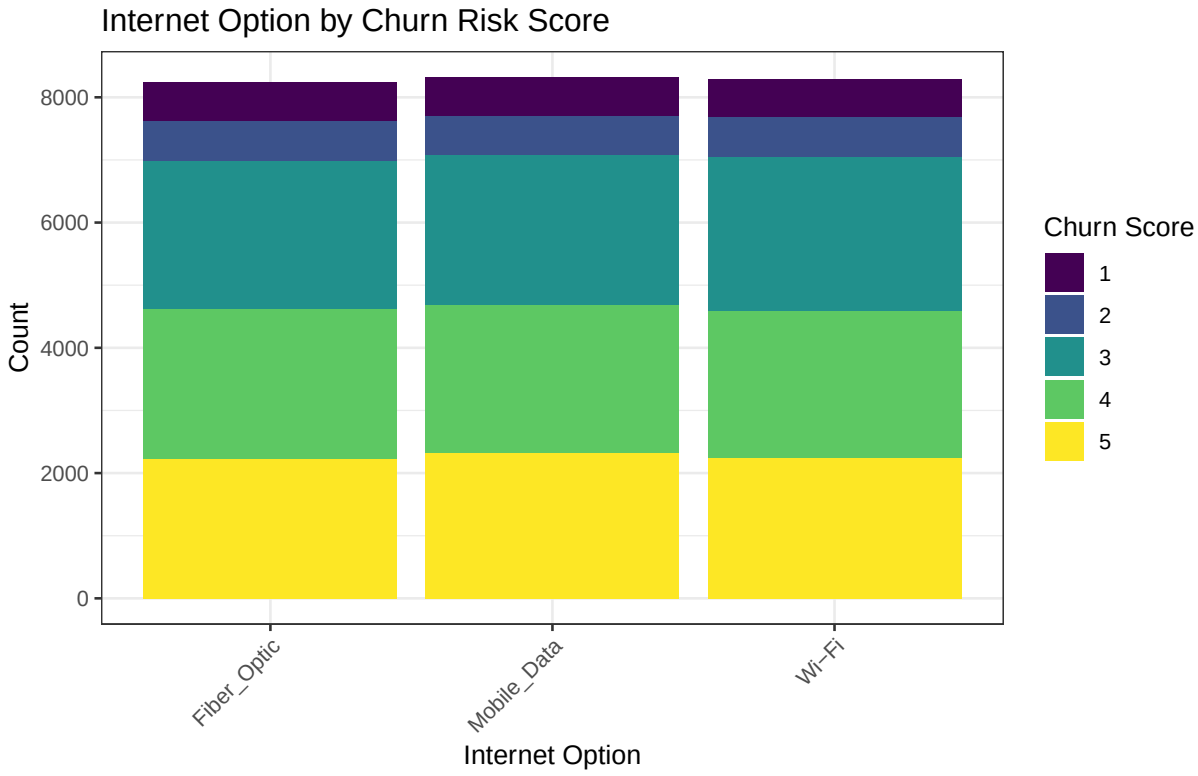
To better understand the relationship between user behavior and churn risk, we included a variety of plots.

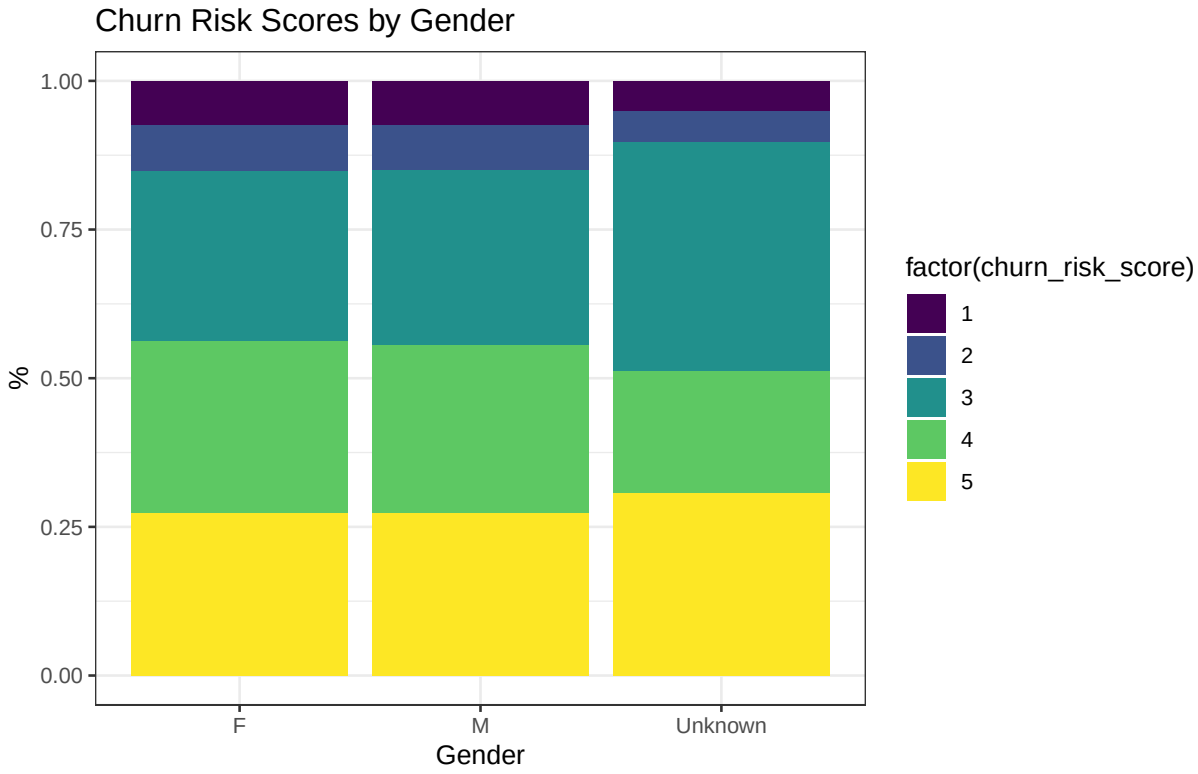

Churn Risk Score Distribution
Count
6000
4000
2000
0
1
2
3
4
5
Churn Risk Score
factor(churn_risk_score)
1
2
3
4
5

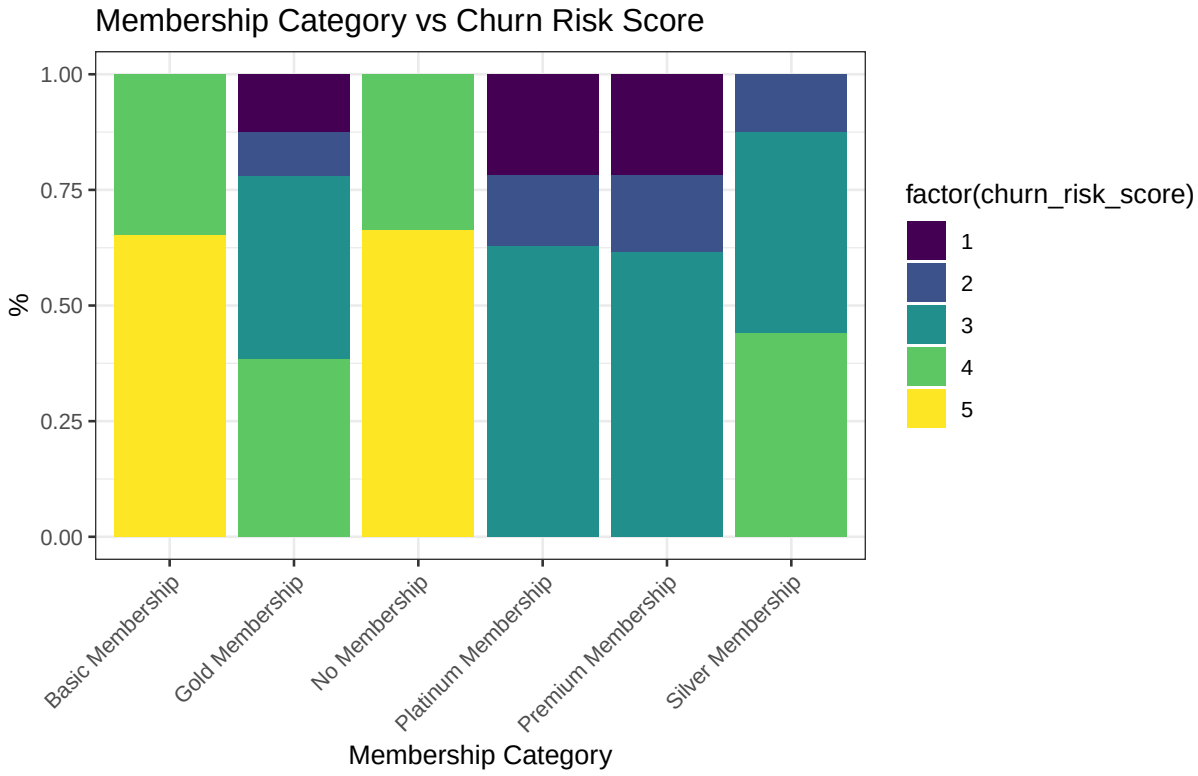

Membership Category Distribution
0
5000
10000
15000
20000
x
n
membership_category
Basic Membership
Gold Membership
No Membership
Platinum Membership
Premium Membership
Silver Membership

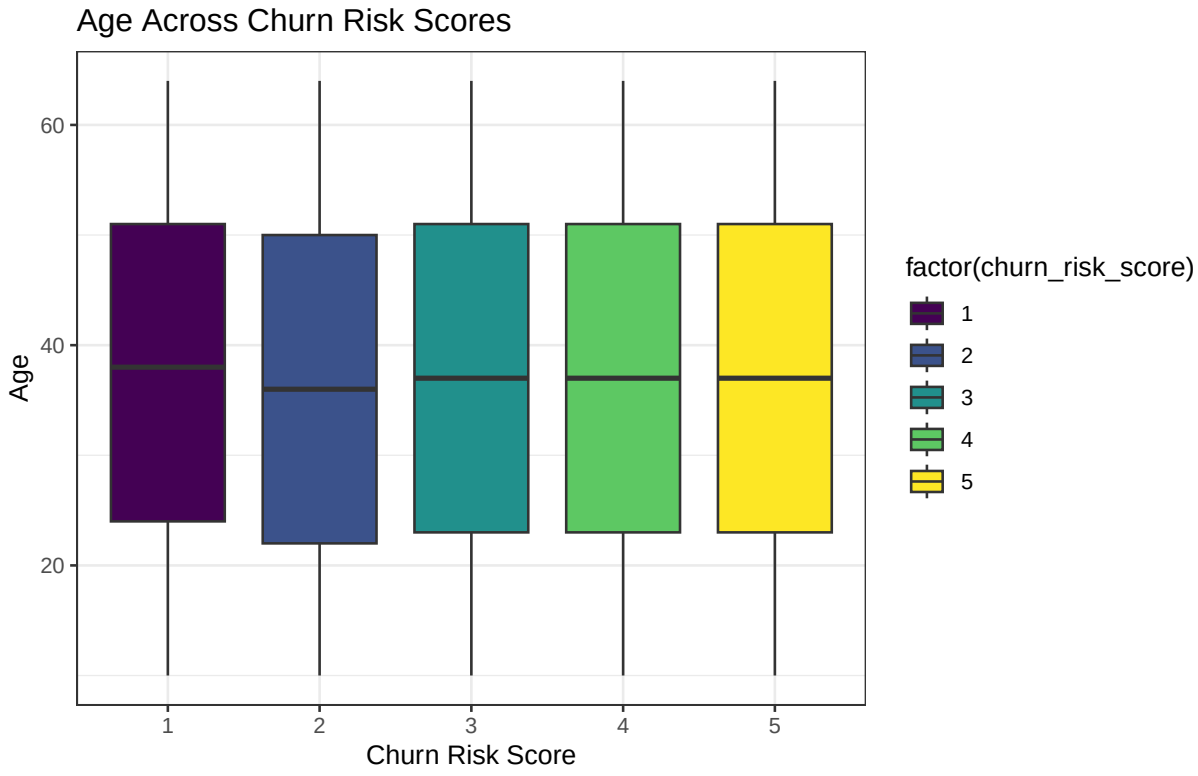

Age Distribution
Count
750
500
250
0
20
40
60
Age

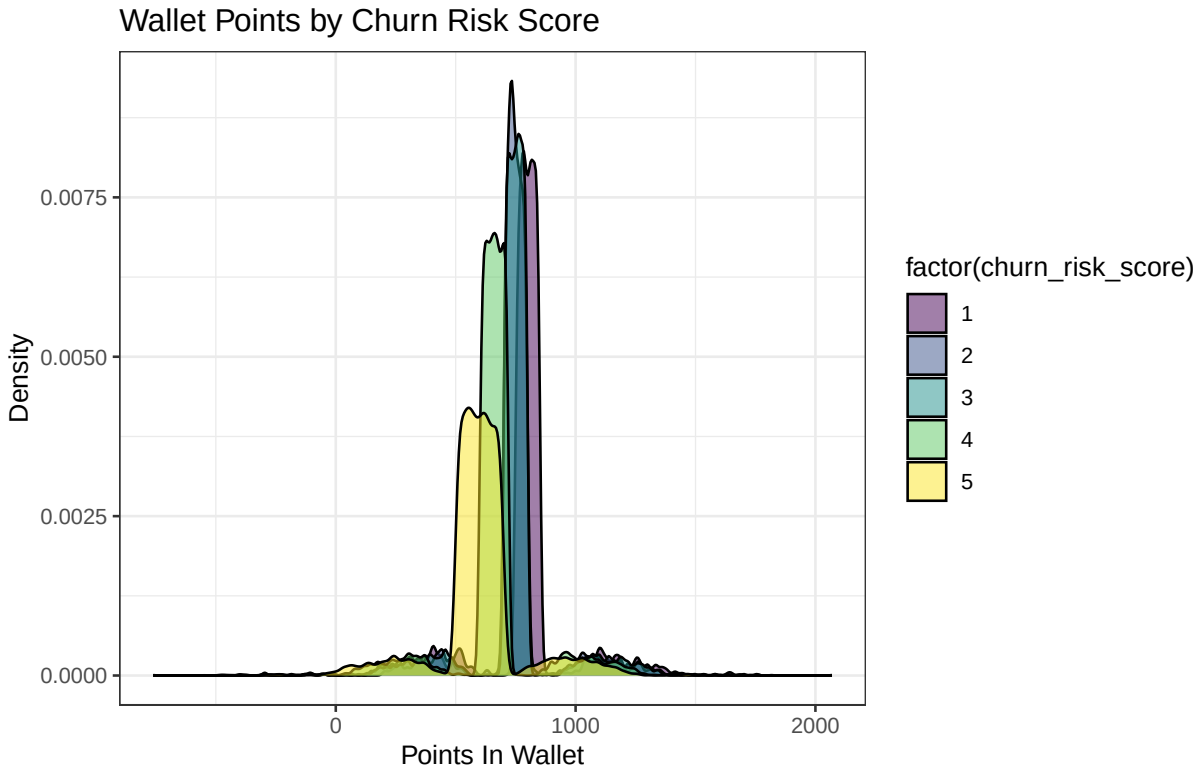

Density of Avg Frequency of Login Days
Density
0.04
0.03
0.02
0.01
0.00
−25
0
25
50
75
Avg Frequency Login Days

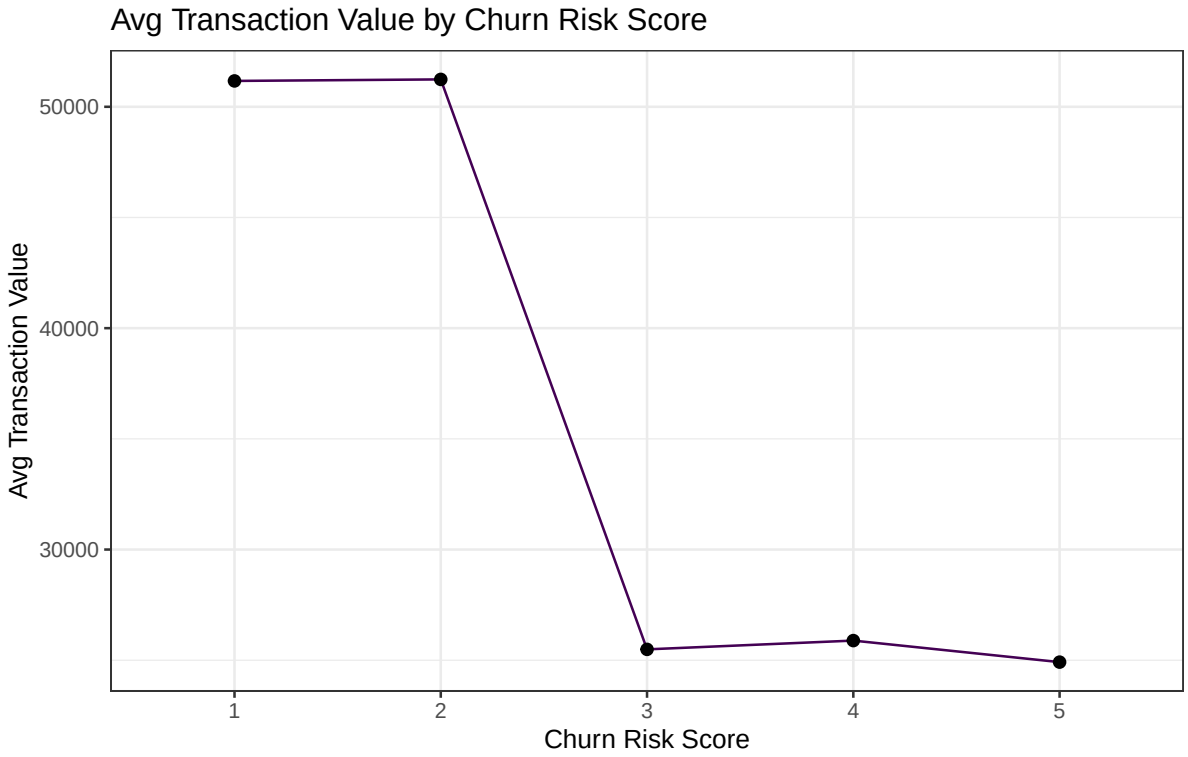

Internet Option by Churn Risk Score
Count
8000
6000
4000
2000
0
Fiber_Optic
Mobile_Data
Wi-Fi
Internet Option
Churn Score
1
2
3
4
5

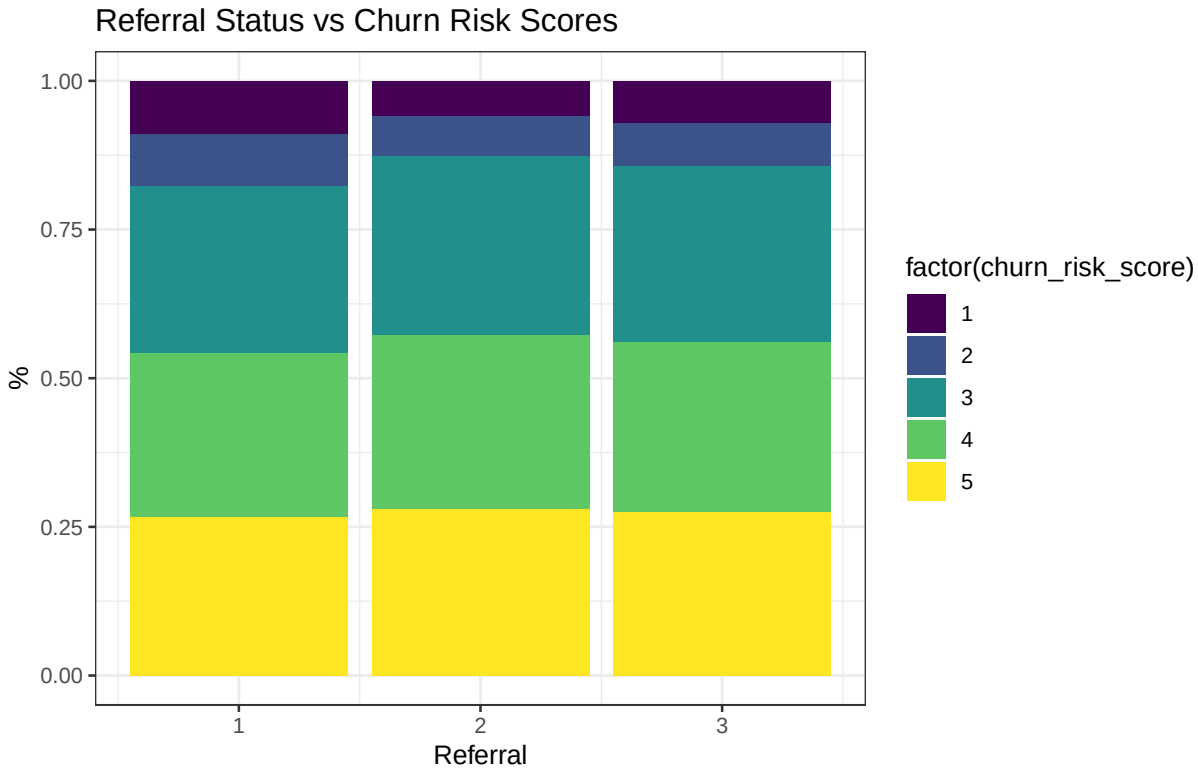

Churn Risk Scores by Gender
1.00
0.75
0.50
0.25
0.00
%
F
M
Unknown
Gender
factor(churn_risk_score)
1
2
3
4
5


Membership Category vs Churn Risk Score
1.00
0.75
0.50
0.25
0.00
%
Basic Membership
Gold Membership
No Membership
Platinum Membership
Premium Membership
Silver Membership
Membership Category
factor(churn_risk_score)
1
2
3
4
5


Age Across Churn Risk Scores
Age
60
40
20
1
2
3
4
5
Churn Risk Score
factor(churn_risk_score)
1
2
3
4
5


Wallet Points by Churn Risk Score
Density
0.0075
0.0050
0.0025
0.0000
0
1000
2000
Points In Wallet
factor(churn_risk_score)
1
2
3
4
5


Avg Transaction Value by Churn Risk Score
Avg Transaction Value
50000
40000
30000
1
2
3
4
5
Churn Risk Score



Modeling Methods

Linear Regression for Average Transaction Value

We implemented three linear regression models to predict each customer's average transaction value using different sets of predictors, including demographic characteristics, engagement behaviors, and their interactions. Model performance was evaluated using 10-fold cross-validation with mean absolute error (MAE) as the primary metric.

Model 1 includes basic demographic and membership info to see if certain types of customers are more likely to churn.

```
model_1 <- lm_spec %>% fit(avg_transaction_value ~ age + gender +
 ↪  region_category + membership_category, data = data)
```

Model 2 focuses on recent customer behavior, since more active users may have lower churn risk.

```
model_2 <- lm_spec %>% fit(avg_transaction_value ~ avg_time_spent +
 ↪  avg_frequency_login_days + days_since_last_login + points_in_wallet, data
 ↪  = data)
```

Model 3 combines features from all domains and includes interaction between age and gender.

```
model_3 <- lm_spec %>% fit(avg_transaction_value ~ age*gender +
 ↪  avg_time_spent + membership_category + past_complaint + internet_option +
 ↪  points_in_wallet, data = data)
```

```
mae_1_in
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 mae     standard      15093.
```

```
mae_2_in
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 mae     standard      15139.
```

```
mae_3_in
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 mae     standard      15089.
```

```
# 10-fold CV MAE
cv_1 <- model_1_cv %>% collect_metrics() %>% filter(.metric == "mae")
cv_2 <- model_2_cv %>% collect_metrics() %>% filter(.metric == "mae")
cv_3 <- model_3_cv %>% collect_metrics() %>% filter(.metric == "mae")
```

```
# 10-fold CV MAE
cv_1
```

```
# A tibble: 1 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 mae     standard   15100.    10    55.2 Preprocessor1_Model1
```

```
cv_2
```

```
# A tibble: 1 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 mae     standard   15142.    10    83.5 Preprocessor1_Model1
```

```
cv_3
```

```
# A tibble: 1 x 6
  .metric .estimator   mean     n std_err .config
  <chr>   <chr>       <dbl> <int>   <dbl> <chr>
1 mae     standard   15099.    10    81.9 Preprocessor1_Model1
```

Model	IN-SAMPLE MAE	10-fold CV MAE
model_1	15092.84	15100.21
model_2	15138.97	15141.82
model_3	15088.64	15098.81

We selected the following linear model based on the lowest cross-validation error:

$$\begin{aligned}\text{avg\_transaction\_value}_i = \beta_0 &+ \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{gender}_i + \beta_3 \cdot (\text{age}_i \times \text{gender}_i) \\ &+ \beta_4 \cdot \text{avg\_time\_spent}_i + \beta_5 \cdot \text{membership\_category}_i + \beta_6 \cdot \text{past\_complaint}_i \\ &+ \beta_7 \cdot \text{internet\_option}_i + \beta_8 \cdot \text{points\_in\_wallet}_i + \varepsilon_i\end{aligned}$$

Table 2: Estimated Coefficients for Final Linear Model

term	estimate	std.error	statistic	p.value
(Intercept)	23154.20927	679.00320	34.10029	0.00000
age	-3.30417	10.72318	-0.30813	0.75798
genderM	39.89084	614.04528	0.06496	0.94820
genderUnknown	-2327.08737	8235.67046	-0.28256	0.77751
avg_time_spent	0.89741	0.30396	2.95244	0.00316
membership_categoryGold Membership	5499.32097	394.44063	13.94208	0.00000
membership_categoryNo Membership	62.73415	375.28374	0.16716	0.86724
membership_categoryPlatinum Membership	9694.10077	451.36484	21.47731	0.00000
membership_categoryPremium Membership	9668.61909	447.52127	21.60483	0.00000
membership_categorySilver Membership	3170.06113	404.17922	7.84321	0.00000
past_complaintYes	-285.87934	241.95659	-1.18153	0.23740
internet_optionMobile_Data	-7.60142	296.40771	-0.02565	0.97954
internet_optionWi-Fi	-279.77984	296.66931	-0.94307	0.34565
points_in_wallet	3.48218	0.64709	5.38130	0.00000
age:genderM	3.19497	15.21981	0.20992	0.83373
age:genderUnknown	38.63170	188.91160	0.20450	0.83797

```
model_3_coefs <- tidy(model_3)

library(kableExtra)

model_3_coefs %>%
  kable(digits = 5, caption = "Estimated Coefficients for Final Linear
   ↪  Model") %>%
  kable_styling(full_width = FALSE)
```

Among all predictors, average time spent on the platform and points in wallet were significantly and positively associated with average transaction value. Specifically, each additional unit of time spent is associated with an increase of approximately $0.90, and each additional point in the wallet corresponds to an increase of $3.48 in transaction value.

Membership category was also a strong predictor. Compared to the baseline group (Basic Membership), customers with Gold, Platinum, Premium, and Silver memberships had significantly higher average transaction values, with coefficients ranging from approximately $3,170 (Silver) to $9,694 (Platinum).

In contrast, demographic variables such as age, gender, and their interaction terms were not statistically significant. Behavioral and engagement factors might be more informative predictors of customer spending than demographics in this context.

Logistic Regression for High Churn Risk

We model the churn risk e using the following predictors:

- `age`
- `gender`
- `points_in_wallet`
- `avg_time_spent`
- `membership_category`

We model the log-odds of being high churn risk ($Y = 1$) as:

$$\log\left(\frac{\mathbb{P}(Y=1 \mid X_1,\ldots,X_5)}{1-\mathbb{P}(Y=1 \mid X_1,\ldots,X_5)}\right) = \beta_0+\beta_1\cdot\text{age}+\beta_2\cdot\text{gender}+\beta_3\cdot\text{points\_in\_wallet}+\beta_4\cdot\text{avg\_time\_spent}+\beta_5\cdot\text{memb}$$

```
logit_model <- glm(
  churn_high ~ age + gender + points_in_wallet + avg_time_spent +
↪  membership_category, data = data, family = binomial(link = "logit")
)
```

```
summary(logit_model)
```

```
Call:
glm(formula = churn_high ~ age + gender + points_in_wallet +
    avg_time_spent + membership_category, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.51978  -0.00005   0.00004   0.00005   2.61243

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        2.280e+01  2.403e+02   0.095    0.924
age                                2.242e-03  1.414e-03   1.586    0.113
genderM                           -6.778e-02  4.538e-02  -1.494    0.135
genderUnknown                     -1.009e+00  7.886e-01  -1.279    0.201
points_in_wallet                  -3.336e-03  1.552e-04 -21.493   <2e-16
```

```
avg_time_spent                           -2.809e-05  5.744e-05  -0.489    0.625
membership_categoryGold Membership       -2.093e+01  2.403e+02  -0.087    0.931
membership_categoryNo Membership         -2.234e-02  3.391e+02   0.000    1.000
membership_categoryPlatinum Membership -4.103e+01  4.009e+02  -0.102    0.918
membership_categoryPremium Membership  -4.106e+01  3.972e+02  -0.103    0.918
membership_categorySilver Membership   -2.071e+01  2.403e+02  -0.086    0.931

(Intercept)
age
genderM
genderUnknown
points_in_wallet                         ***
avg_time_spent
membership_categoryGold Membership
membership_categoryNo Membership
membership_categoryPlatinum Membership
membership_categoryPremium Membership
membership_categorySilver Membership
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34118  on 24852  degrees of freedom
Residual deviance: 11056  on 24842  degrees of freedom
AIC: 11078

Number of Fisher Scoring iterations: 19
```

Based on the results, we can observe that from this data context:

- age: For each 1-year increase in age, the odds of high churn risk increase slightly.
- genderMale: Being male is associated with slightly lower odds of high churn risk compared to being female.
- genderUnknown: Having unknown gender is associated with much lower odds of high churn risk compared to being female.
- points_in_wallet: For each additional point in the wallet, the odds of high churn risk decrease slightly.
- avg_time_spent: For each additional unit of average time spent, the odds of high churn risk decrease very slightly.

- Gold Membership: Being a Gold member is associated with very low odds of high churn risk compared to Basic.
- No Membership: Having no membership is associated with little to no change in churn risk compared to Basic.
- Platinum Membership: Being a Platinum member is associated with very low odds of high churn risk compared to Basic.
- Premium Membership: Being a Premium member is associated with very low odds of high churn risk compared to Basic.
- Silver Membership: Being a Silver member is associated with very low odds of high churn risk compared to Basic.

Since **points in wallet** appeared to be a significant predictor of churn risk, we provide a more detailed interpretation of its effect. The estimated coefficient was

$$\hat{\beta}_{\text{points\_in\_wallet}} = -0.003336$$

```
exp(-0.003336)
```

```
[1] 0.9966696
```

We found that each additional point in the wallet **reduces the odds of being high churn risk by about 0.33%**. We also simulated predictions for representative customers from our dataset.

```
new_data <- data.frame(
  age = c(25, 45),
  gender = factor(c("F", "M"), levels = levels(data$gender)),
  points_in_wallet = c(100, 300),
  avg_time_spent = c(20, 5),
  membership_category = factor(c("Basic Membership", "Gold Membership"),
                               levels = levels(data$membership_category))
)

predicted_probs <- predict(logit_model, newdata = new_data, type =
 ↪  "response")

cbind(new_data, predicted_probability = predicted_probs)
```

```
  age gender points_in_wallet avg_time_spent membership_category
1  25      F              100             20    Basic Membership
2  45      M              300              5     Gold Membership
  predicted_probability
1             1.0000000
2             0.7120415
```

From our data context, we could predict that a 25-year-old female with 100 wallet points, 20 average time spent, and Basic Membership has a predicted probability of 1 of being high churn risk. On the other hand, a 45-year-old male with 300 wallet points, 5 average time spent, and Gold Membership has a predicted probability of 0.712 of being high churn risk.

It is worth noting that logistic regression is better suited for binary outcomes because it ensures predicted probabilities stay between 0 and 1, while linear regression can produce invalid probabilities outside that range. Logistic regression models the log-odds, but its coefficients are more difficult to interpret.

Conclusion

This project applied linear regression to predict average transaction value and logistic regression to classify high churn risk. Results indicated that behavioral features, such as time spent on the platform and points in wallet, were more predictive than demographic variables. Cross-validation was used to compare model performance, and the final models highlighted key factors associated with customer spending and retention. Overall, the analysis suggests that more engaged users tend to spend more and are less likely to churn.