Customer Churn Risk Score Prediction

Co-leads: Kevin Gui and Dave Wang
Members: Caroline, Gabriel, Jayden, Emilie, Kyle, Iain, Tyler, and Karissa

Table of contents

01 Problem Statement

Dataset and what to analyze

Introduction to the Churn Problem

Churn is defined as the loss of customers over a period of time for a business. (stop using services, unsubscribing, etc)

It is thus crucial to understand where the churning comes from, their potential issues, and modeling for predicting the churn risk scores.

Understanding the Dataset

- 36992 rows and 25 columns
- 19 categorical variables
- 5 numerical variables

```
Summary of the dataset is
                age  days_since_last_login  avg_time_spent  \
count  36992.000000           36992.000000    36992.000000
mean      37.118161             -41.915576      243.472334
std       15.867412             228.819900      398.289149
min       10.000000            -999.000000    -2814.109110
25%       23.000000               8.000000       60.102500
50%       37.000000              12.000000      161.765000
75%       51.000000              16.000000      356.515000
max       64.000000              26.000000     3235.578521

       avg_transaction_value  points_in_wallet  churn_risk_score
count           36992.000000      33549.000000      36992.000000
mean            29271.194003        686.882199          3.463397
std             19444.806226        194.063624          1.409661
min               800.460000       -760.661236         -1.000000
25%             14177.540000        616.150000          3.000000
50%             27554.485000        697.620000          4.000000
75%             40855.110000        763.950000          5.000000
max             99914.050000       2069.069761          5.000000
```

Categorical and Numerical Variables

	customer_id	Name	age	gender	security_no	region_category	membership_category	joining_date	joined_through_referral
	fffe43004900440036003000030003800	Pattie Morrisey	18	F	XW0DQ7H	Village	Platinum Membership	2017-08-17	No
	fffe43004900440032003100300035003700	Traci Peery	32	F	5K0N3X1	City	Premium Membership	2017-08-28	?
	fffe43004900440031003900320003600	Merideth Mcmeen	44	F	1F2TCL3	Town	No Membership	2016-11-11	Yes
	fffe43004900440036003000330031003600	Eufemia Cardwell	37	M	VJGJ33N	City	No Membership	2016-10-29	Yes
	fffe43004900440031003900350030003600	Meghan Kosak	31	F	SVZXCWB	City	No Membership	2017-09-12	No

	referral_id	preferred_offer_types	medium_of_operation	internet_option	last_visit_time	days_since_last_login	avg_time_spent	avg_transaction_value
	xxxxxxxx	Gift Vouchers/Coupons	?	Wi-Fi	16:08:02	17	300.63	53005.25
	CID21329	Gift Vouchers/Coupons	Desktop	Mobile_Data	12:38:13	16	306.34	12838.38
	CID12313	Gift Vouchers/Coupons	Desktop	Wi-Fi	22:53:21	14	516.16	21027.00
	CID3793	Gift Vouchers/Coupons	Desktop	Mobile_Data	15:57:50	11	53.27	25239.56
	xxxxxxxx	Credit/Debit Card Offers	Smartphone	Mobile_Data	15:46:44	20	113.13	24483.66

	avg_frequency_login_days	points_in_wallet	used_special_discount	offer_application_preference	past_complaint	complaint_status	feedback	churn_risk_score
	17.0	781.75	Yes	Yes	No	Not Applicable	Products always in Stock	2
	10.0	NaN	Yes	No	Yes	Solved	Quality Customer Care	1
	22.0	500.69	No	Yes	Yes	Solved in Follow-up	Poor Website	5
	6.0	567.66	No	Yes	Yes	Unsolved	Poor Website	5
	16.0	663.06	No	Yes	Yes	Solved	Poor Website	5

02

Processing

How is Churn Score interpreted?

Low Risk

Churn scores of 1-2 show that there is a minimal risk of a customer leaving the business.

Medium Risk

Churn scores of 3 show that there is a moderate risk of a customer leaving if not engaged with properly.

High Risk

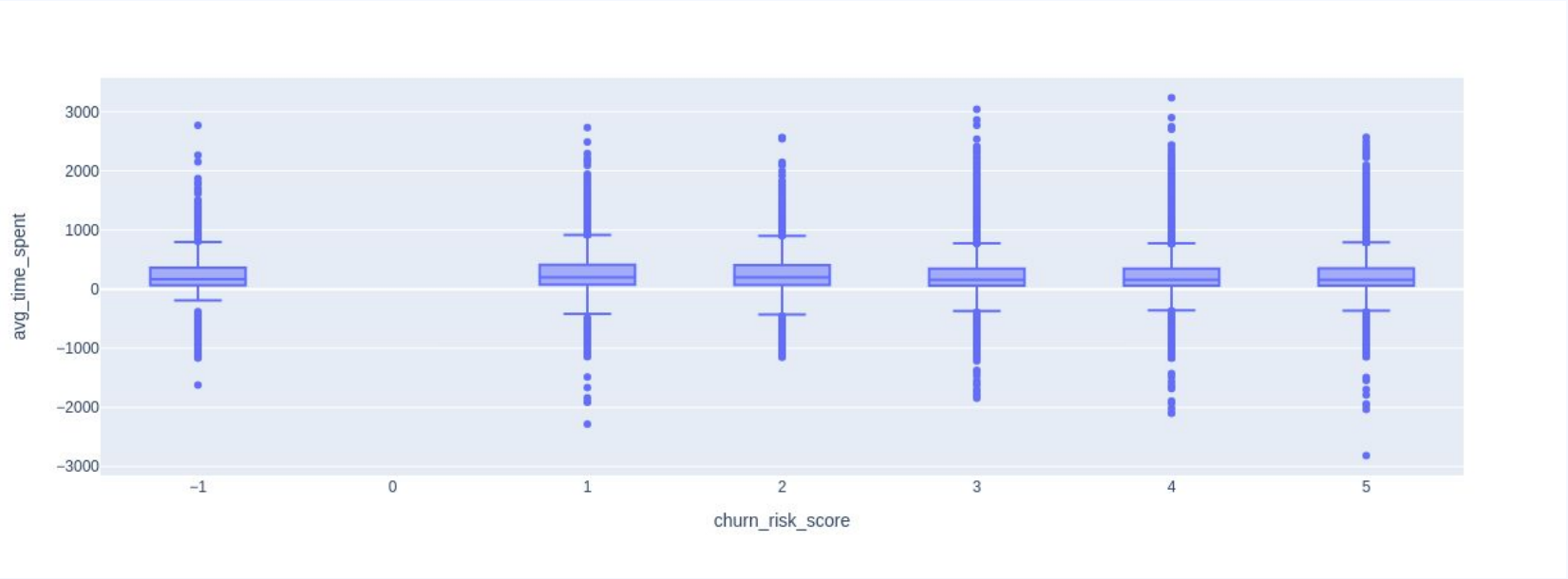
Churn scores of 4-5 show a high amount of customer dissatisfaction, with a high risk of customers leaving.

EDA

Membership Category vs Churn Risk Score

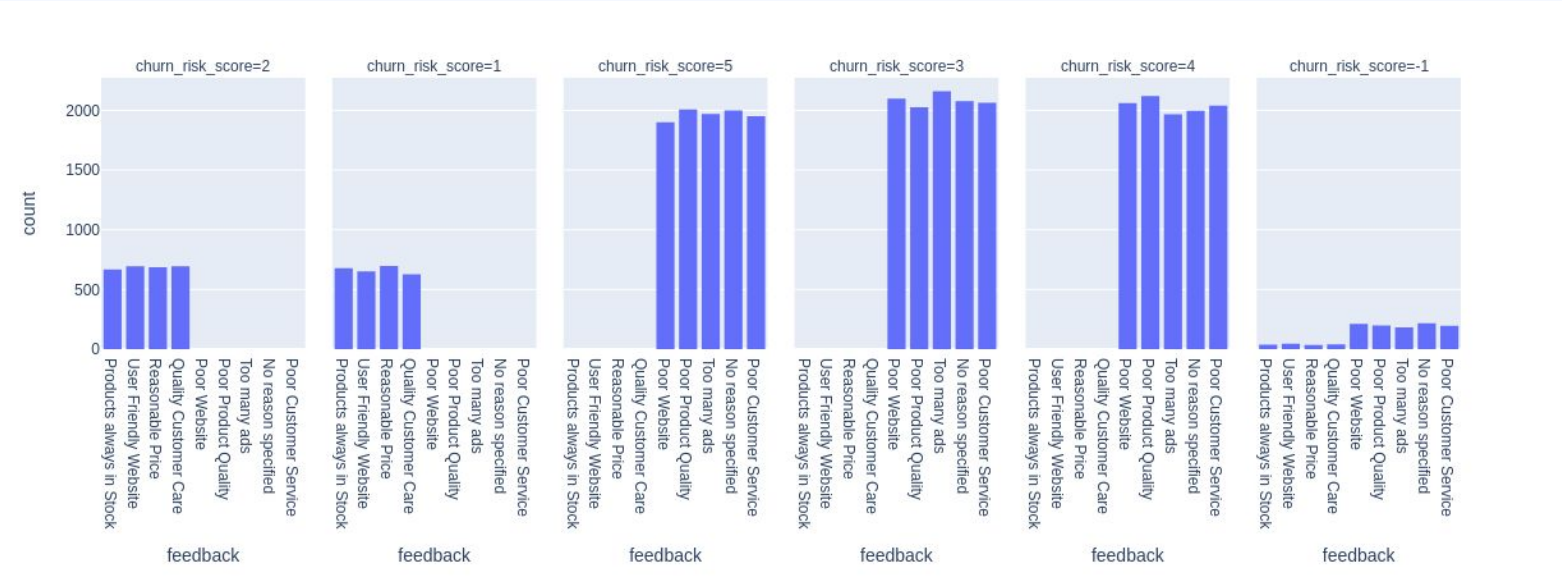
Basic/No membership average transaction value is capped at 50k, those with higher churn scores typically had little to no membership.

Average Time Spent vs Churn Risk Score



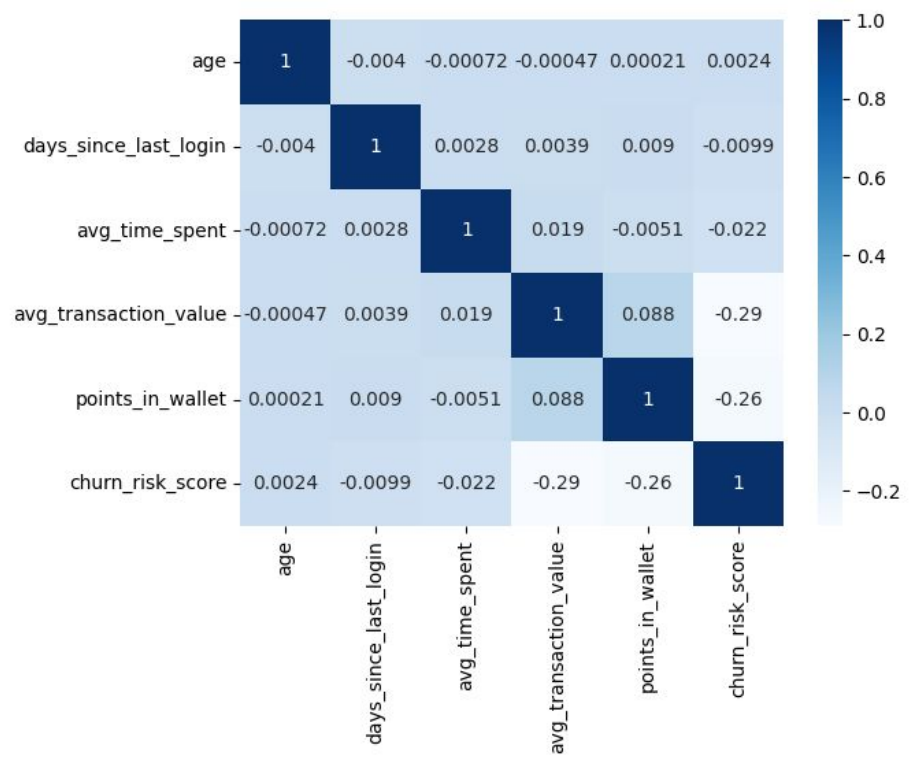
Box plots with higher churn risk score had larger variances of average time spent.

Feedback Vs. Churn Risk Score



Those with positive feedback had lower churn risk score, those with negative feedback had higher churn risk score.

Correlation Matrix

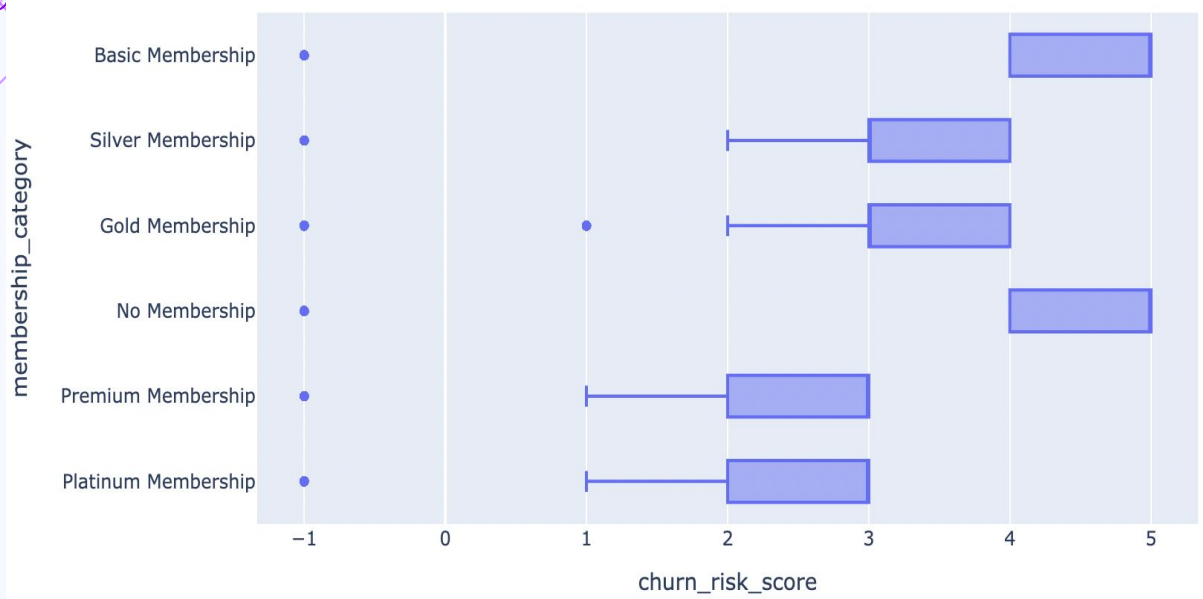


Correlation between each numerical variable & with churn_risk_score

Indicates stronger (and negative) linear relationship between:

- Avg_transaction_value & churn_risk_score
- Points_in_wallet & churn_risk_score

Membership Category vs Churn Risk Score

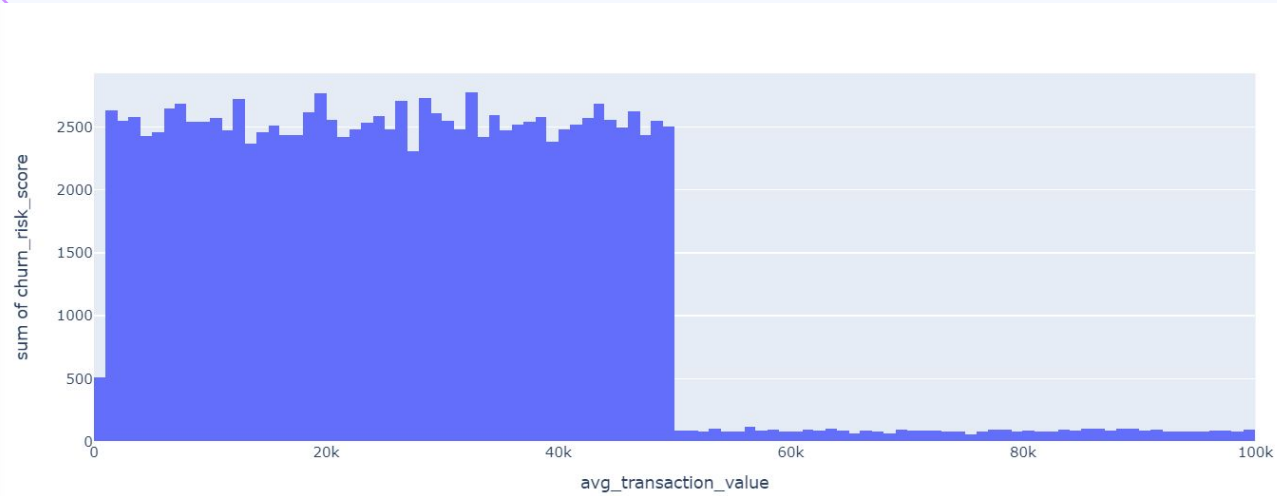


Gauge the churn risk score based off of the membership category.

Check if there are any outliers

See the IQR of the boxplot

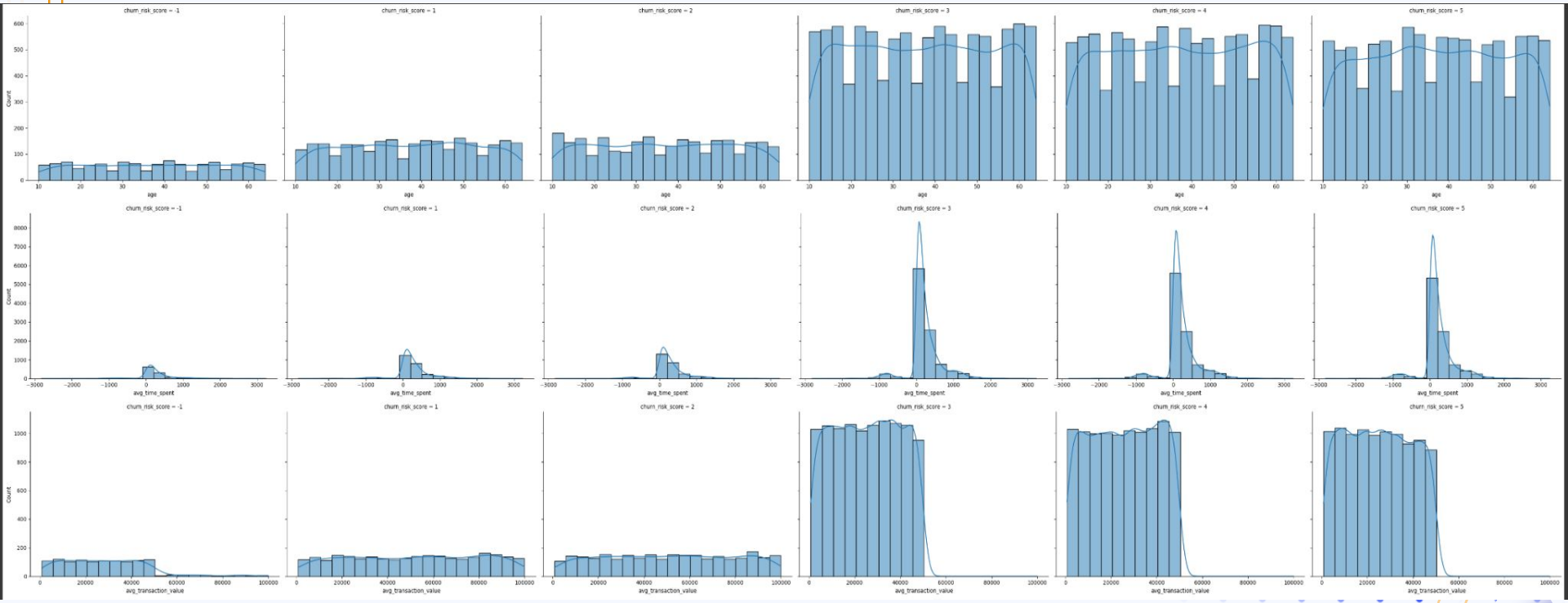
Average Transaction Value vs Total Churn



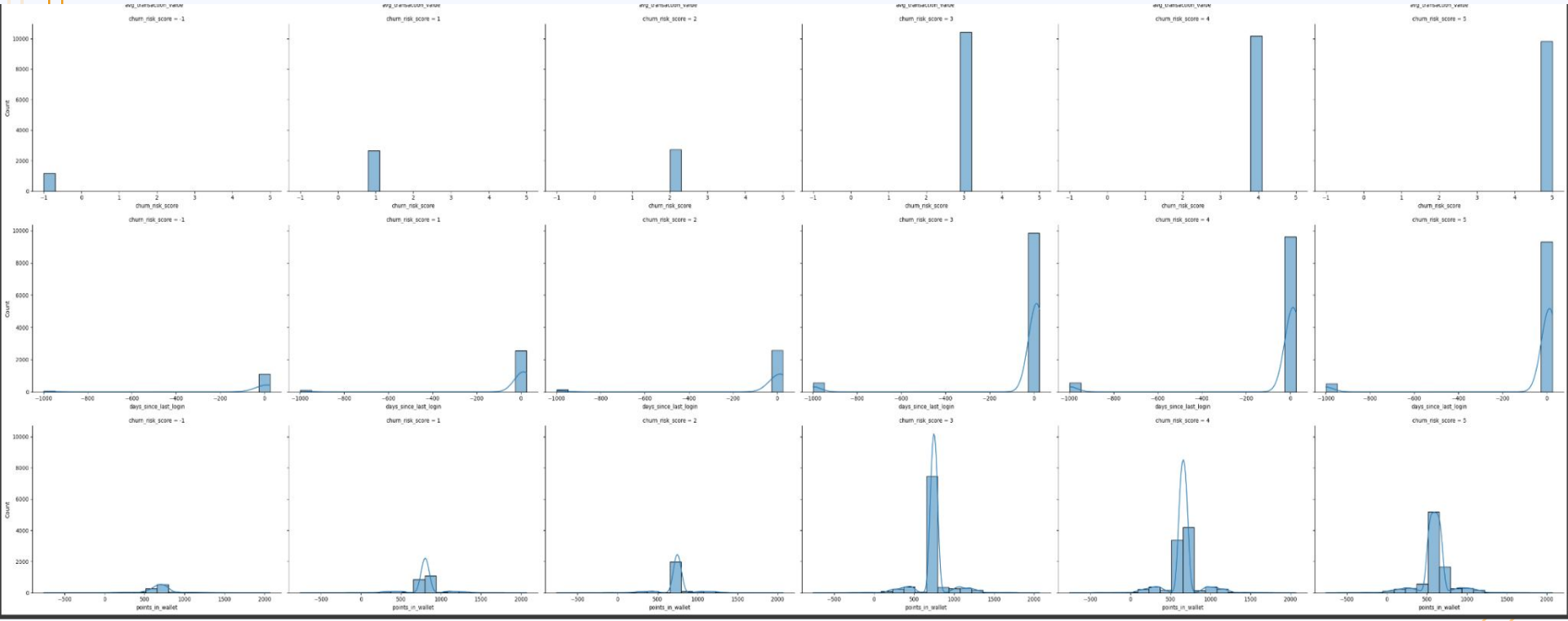
Shows the total amount of churn for different amounts of average transaction values.

At higher transaction values, there is lower overall churn risk.

Distributions



Distributions



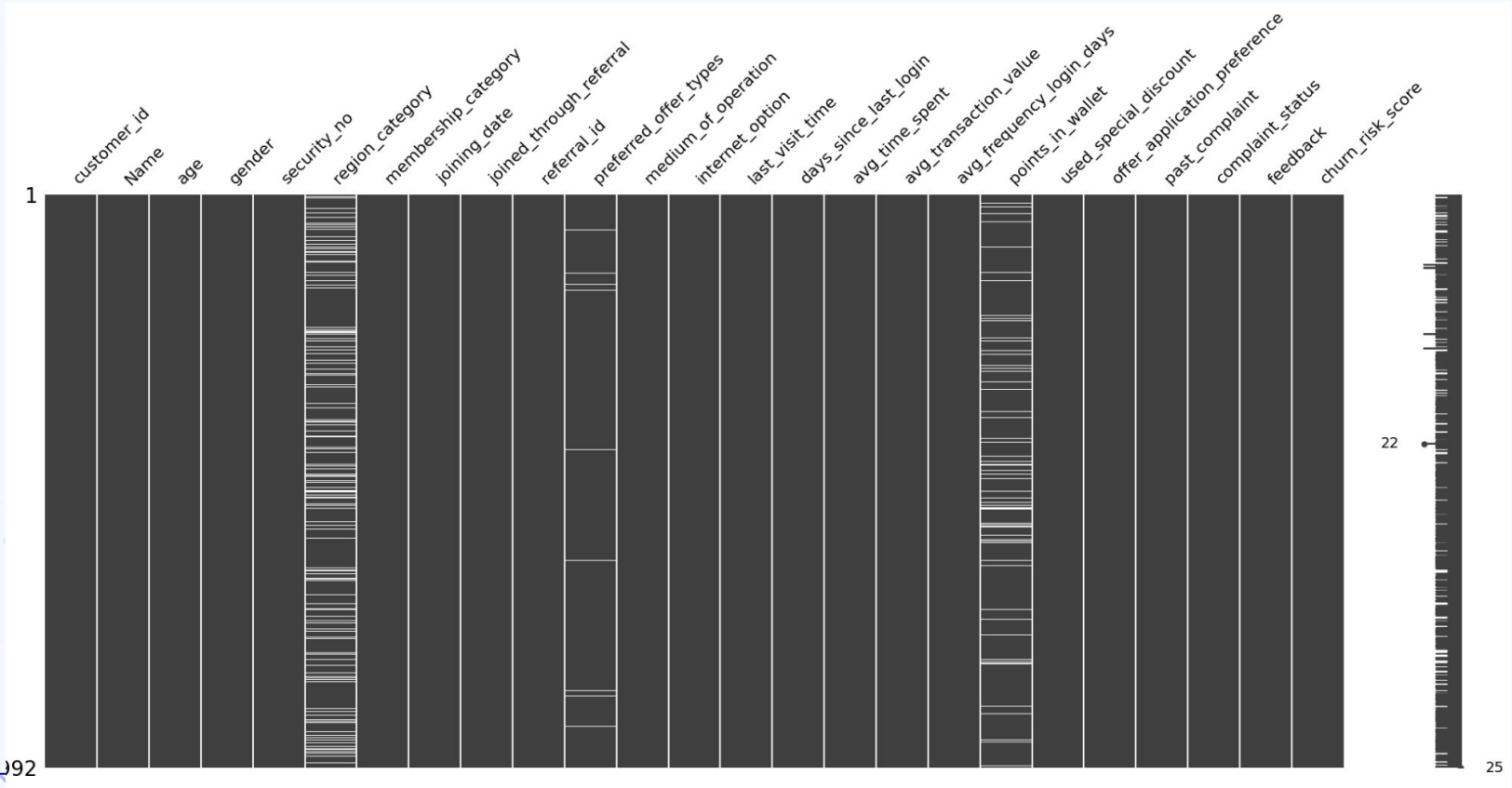
Missing Values

Data Cleaning

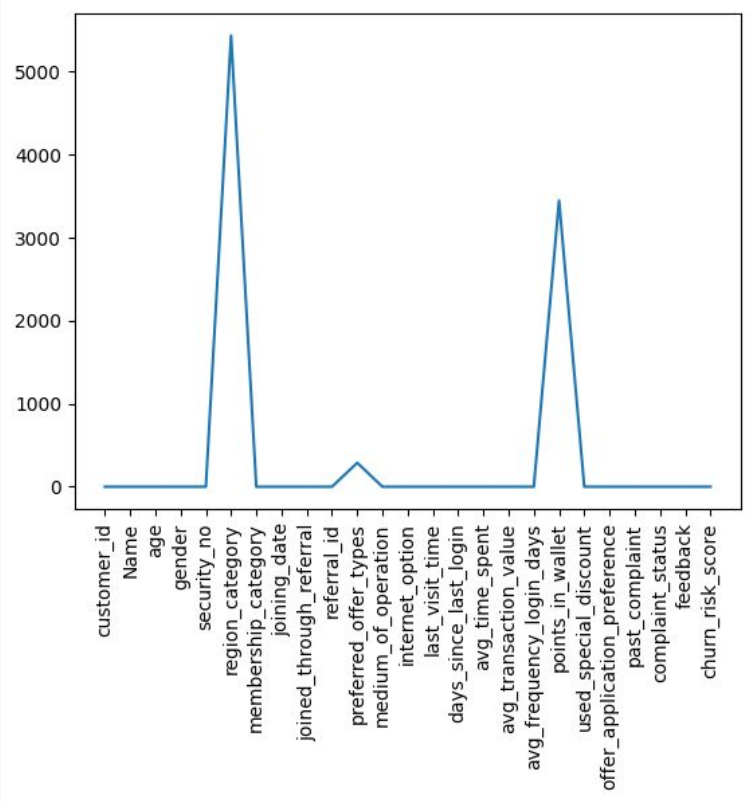
joining_date	joined_through_referral	referral_id
2017-08-17	No	xxxxxxxx
2017-08-28	?	CID21329
2016-11-11	Yes	CID12313
2016-10-29	Yes	CID3793
2017-09-12	No	xxxxxxxx

joining_date	joined_through_referral	referral_id
20170817	No	NaN
20170828	unknown	CID21329
20161111	Yes	CID12313
20161029	Yes	CID3793
20170912	No	NaN

NA Values Across All Columns



NA Values Across All Columns



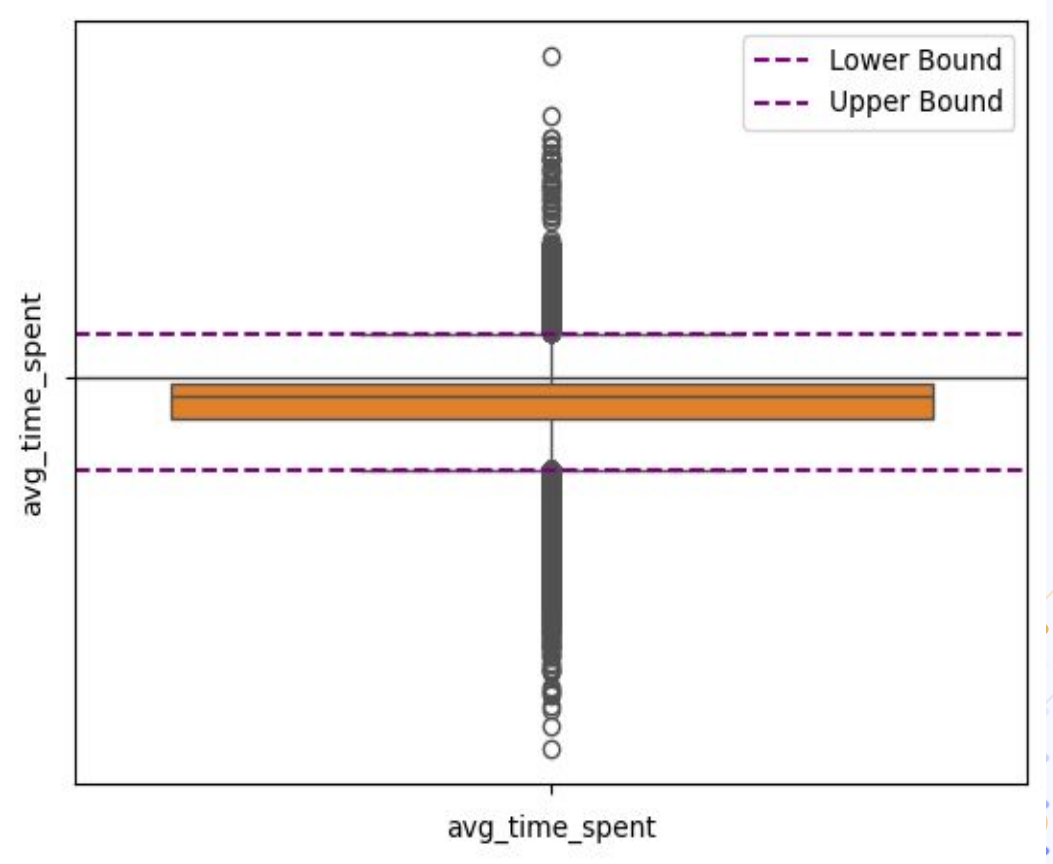

```
customer_id                     0
Name                            0
age                             0
gender                          0
security_no                     0
region_category              5428
membership_category             0
joining_date                    0
joined_through_referral         0
referral_id                     0
preferred_offer_types         288
medium_of_operation             0
internet_option                 0
last_visit_time                 0
days_since_last_login           0
avg_time_spent                  0
avg_transaction_value           0
avg_frequency_login_days        0
points_in_wallet             3443
used_special_discount           0
offer_application_preference    0
past_complaint                  0
complaint_status                0
feedback                        0
churn_risk_score                0
dtype: int64
```

Outliers

To prevent from a few extreme values potentially skewing the model performance, we identify those

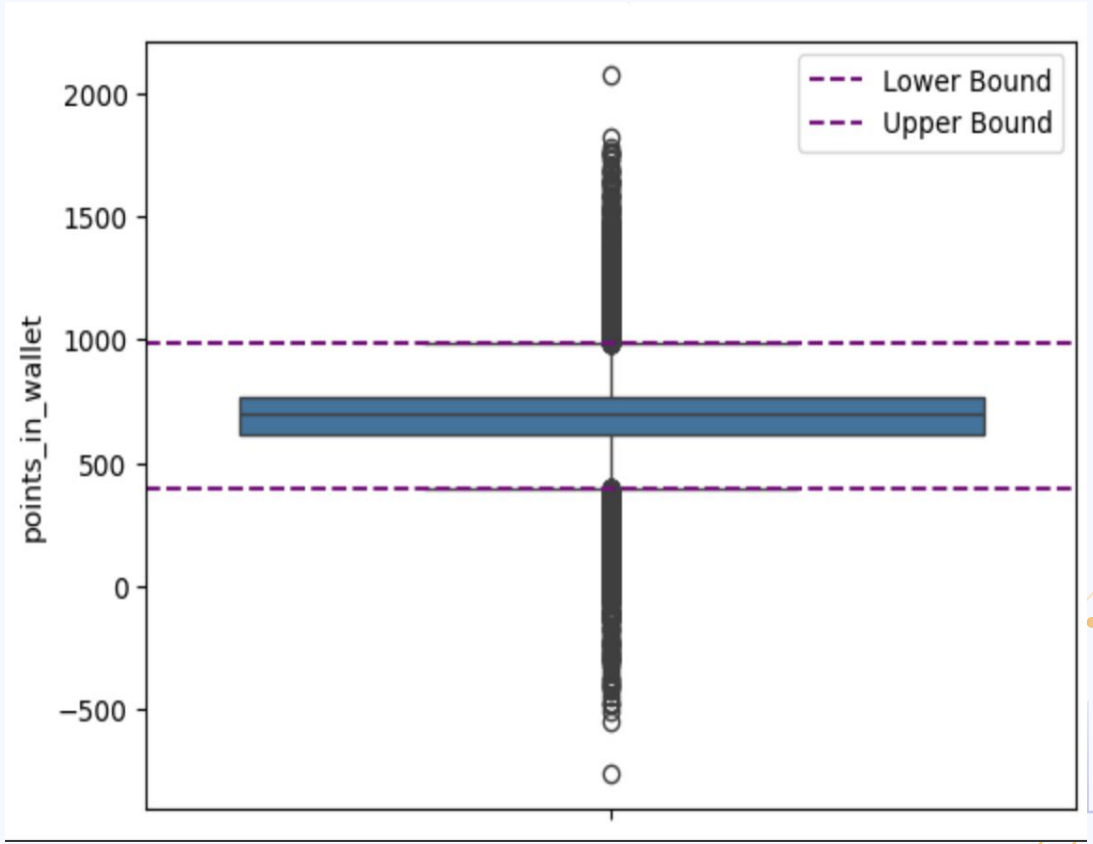
Winsorize Outliers

- Boxplot to detect numerical values beyond IQR
- Values outside IQR → clipped to upper & lower bounds
- See the purple line?

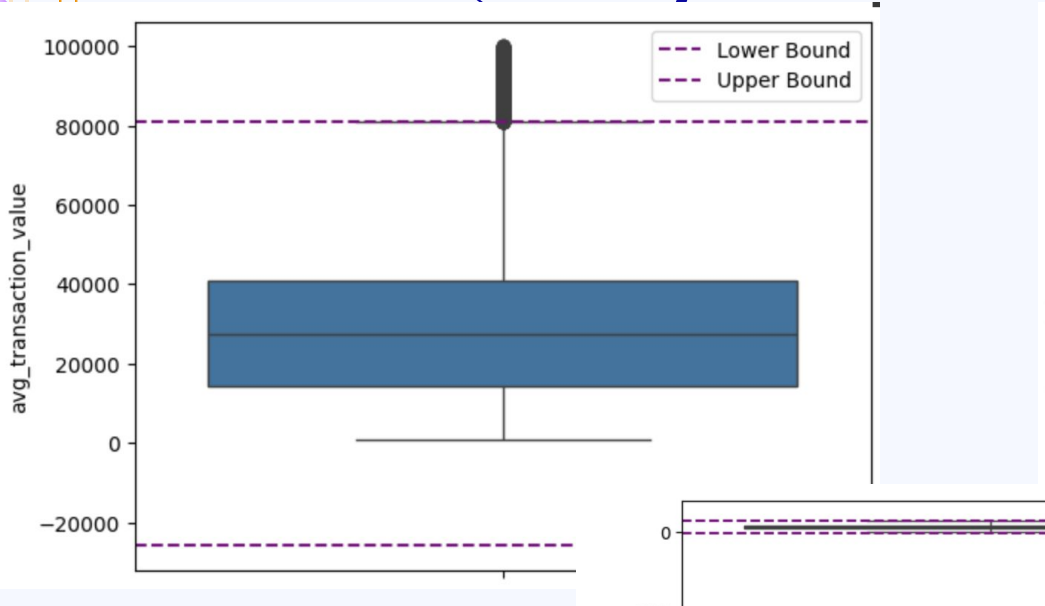


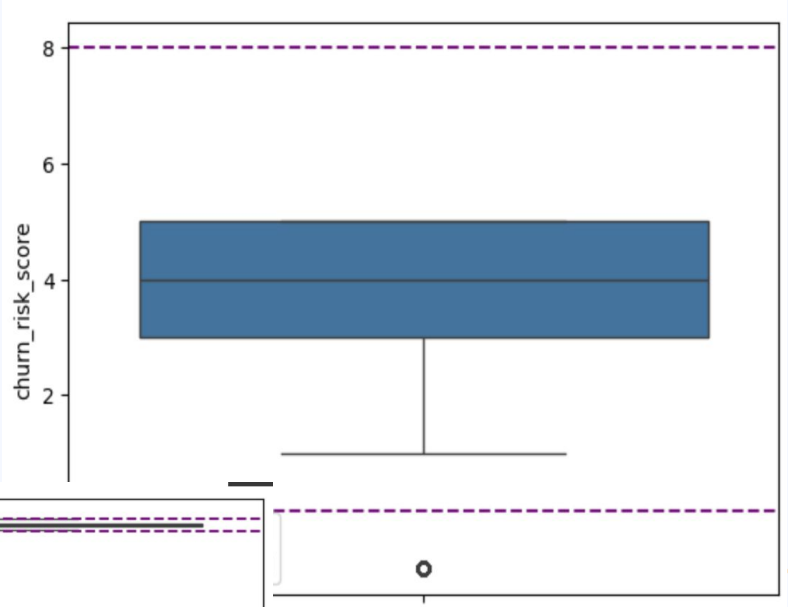
Winsorize Outliers

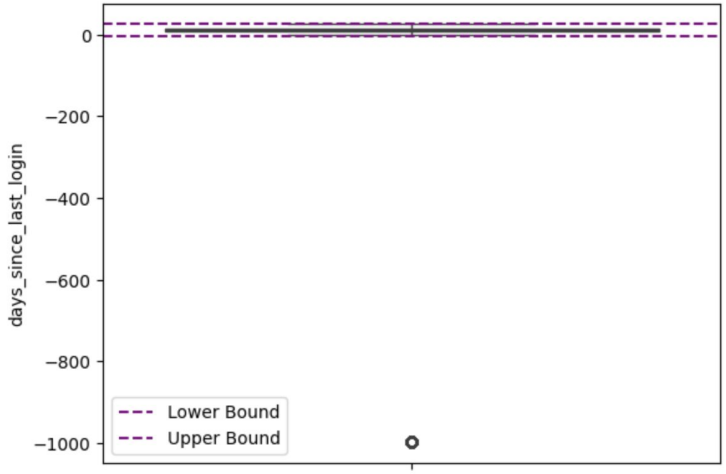
- Boxplot to detect numerical values beyond IQR
- Values outside IQR → clipped to upper & lower bounds



Fewer/No Outliers







03

Modeling

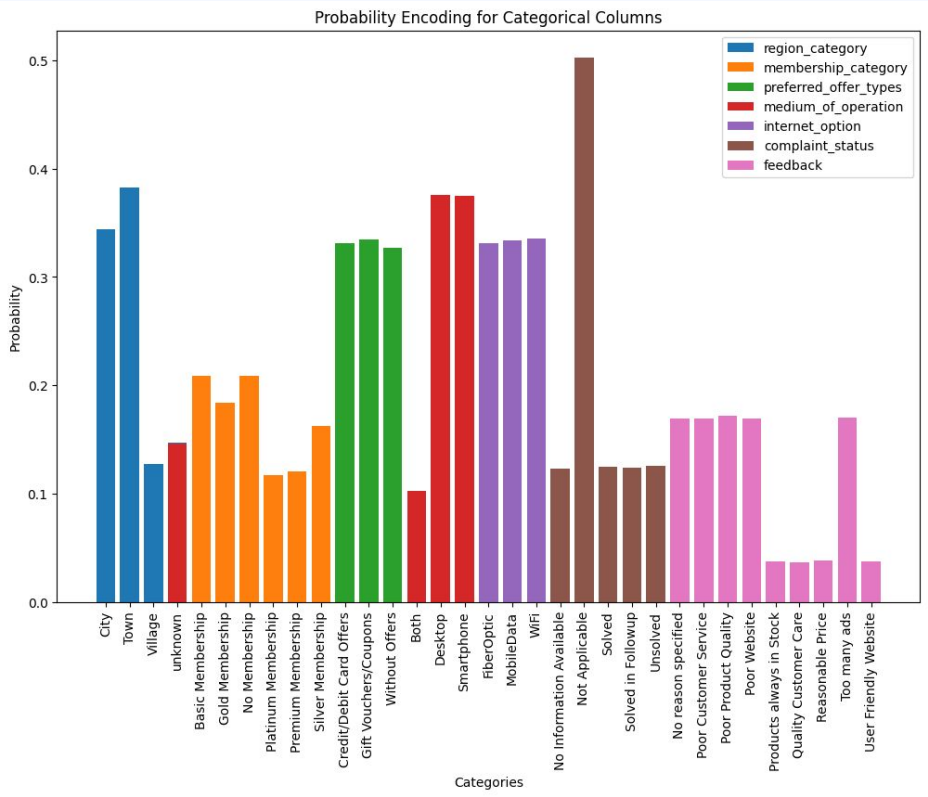
FE & Encoding

Frequency Encoding

- For each categorical variable, shows counts of each subcategory within that category
- We included “unknown” as a possible subcategory


```
{
  "region_category": {
    "City": 12315,
    "Town": 13702,
    "Village": 4549,
    "unknown": 5263
  },
  "membership_category": {
    "Basic Membership": 7473,
    "Gold Membership": 6574,
    "No Membership": 7466,
    "Platinum Membership": 4202,
    "Premium Membership": 4308,
    "Silver Membership": 5806
  },
  "preferred_offer_types": {
    "Credit/Debit Card Offers": 11860,
    "Gift Vouchers/Coupons": 11977,
    "Without Offers": 11716,
    "unknown": 276
  },
```

Frequency/Probability Encoding




```
{'region_category': {'City': 0.34372,
  'Town': 0.38243,
  'Village': 0.12696,
  'unknown': 0.14689},
 'membership_category': {'Basic Membership': 0.20857,
  'Gold Membership': 0.18348,
  'No Membership': 0.20838,
  'Platinum Membership': 0.11728,
  'Premium Membership': 0.12024,
  'Silver Membership': 0.16205},
 'preferred_offer_types': {'Credit/Debit Card Offers': 0.33102,
  'Gift Vouchers/Coupons': 0.33428,
  'Without Offers': 0.327,
  'unknown': 0.0077},
 'medium_of_operation': {'Both': 0.10268,
  'Desktop': 0.37612,
  'Smartphone': 0.37523,
  'unknown': 0.14597},
 'internet_option': {'FiberOptic': 0.33102,
  'MobileData': 0.33361,
  'WiFi': 0.33537},
 'complaint_status': {'No Information Available': 0.12311,
  'Not Applicable': 0.50258,
  'Solved': 0.12468,
  'Solved in Followup': 0.12401,
  'Unsolved': 0.12562},
 'feedback': {'No reason specified': 0.16947,
  'Poor Customer Service': 0.16903,
  'Poor Product Quality': 0.1717,
  'Poor Website': 0.16914,
  'Products always in Stock': 0.03754,
  'Quality Customer Care': 0.03684,
```

One Hot Encoding

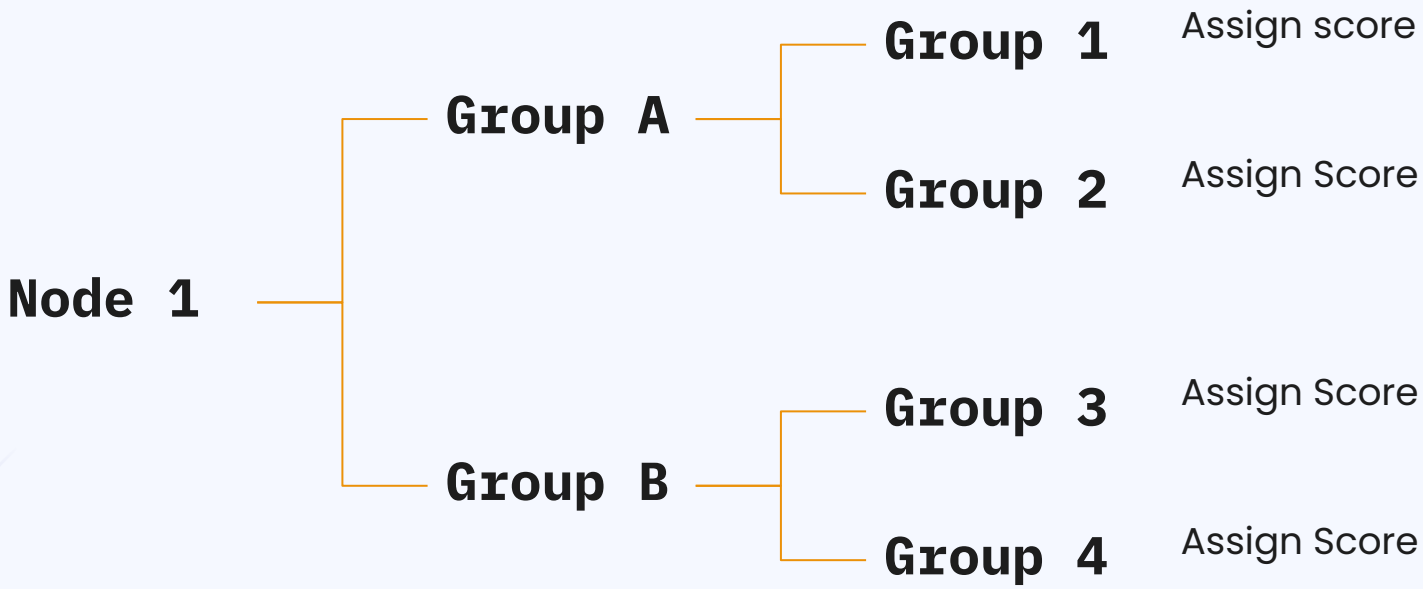
gender_F	gender_M
1	0
0	1
0	1
1	0
0	1

1 - Yes
0 - No

Converts “Yes” or “No” Categorical Variables into binary numbers (1 or 0).

Model

Decision Tree



XGBoost

Build Decision Trees sequentially, instead of 1 single tree

Scalable, accurate, relatively fast

But easy to overfit

1 tree
Calculate Error
Build 1 tree to correct

Optimization + Metrics

Hyper-parameters

Criterion (“gini”, “entropy”)
- Evaluates the quality of a split in the decision tree

Max Depth
- How deep the tree can go

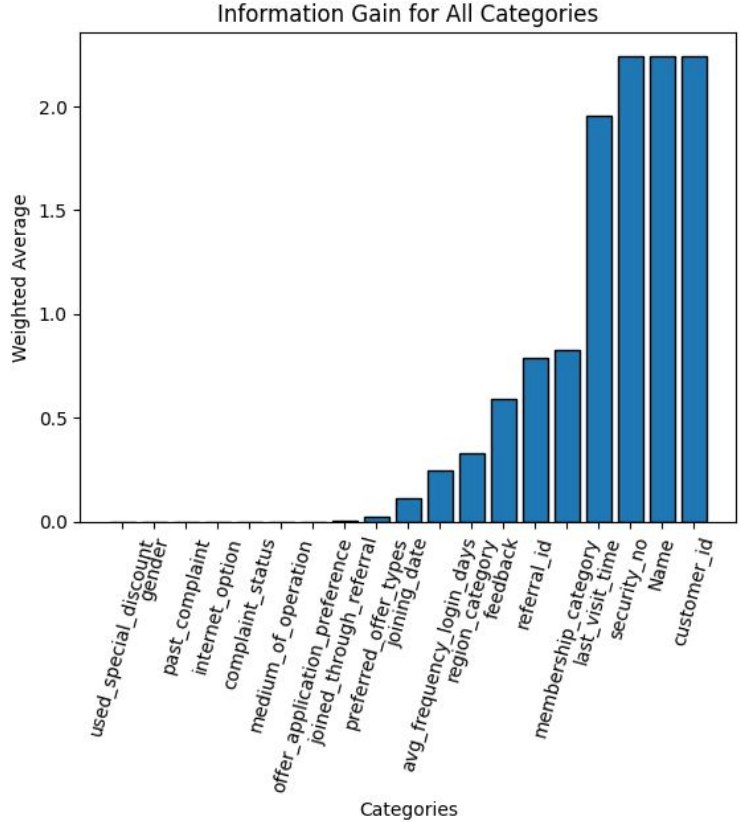
Max Features
- Maximum number of features when looking for the best split

Min Sample Leaf
- Minimum number of samples needed in order to be at a leaf node

Min Samples Split
- Minimum number of samples needed to split an internal node

Understanding Entropy

Information gain as a weighted average of entropy, subtracted from expected entropy.



- Graph tells us how significant the category is to the customer churn score
- Security number, Name, and Customer Id are most significant
- But those columns will be removed, and the next 3 most important: **last_visit_time, membership_category, referral_id**

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Minus the weighted average of the above quantity

	churn_risk_score			
	mean	median	min	max
membership_category				
Basic Membership	4.654757	5.0	4.0	5.0
Gold Membership	3.031183	3.0	1.0	4.0
No Membership	4.660863	5.0	4.0	5.0
Platinum Membership	2.411471	3.0	1.0	3.0
Premium Membership	2.409006	3.0	1.0	3.0
Silver Membership	3.317258	3.0	2.0	4.0

	mean	median	min	max
feedback				
No reason specified	3.986825	4.0	3.0	5.0
Poor Customer Service	3.981506	4.0	3.0	5.0
Poor Product Quality	3.997074	4.0	3.0	5.0
Poor Website	3.967492	4.0	3.0	5.0
Products always in Stock	1.495911	1.0	1.0	2.0
Quality Customer Care	1.525758	2.0	1.0	2.0
Reasonable Price	1.495658	1.0	1.0	2.0
Too many ads	3.968832	4.0	3.0	5.0
User Friendly Website	1.516345	2.0	1.0	2.0

Variables like gender, referrals, discount, region categories do not have significant discrepancies in churn risk score, but variables like membership category and complaint status do.

	churn_risk_score			
	mean	median	min	max
complaint_status				
No Information Available	3.621854	4.0	1.0	5.0
Not Applicable	3.593991	4.0	1.0	5.0
Solved	3.630848	4.0	1.0	5.0
Solved in Followup	3.636057	4.0	1.0	5.0
Unsolved	3.602311	4.0	1.0	5.0

Complaint status seems to have high churn risk across the board.

complaint_status	feedback	
No Information Available	No reason specified	3.959103
	Poor Customer Service	3.991935
	Poor Product Quality	3.956522
	Poor Website	3.975069
	Products always in Stock	1.515337
	Quality Customer Care	1.536424
	Reasonable Price	1.446541
	Too many ads	4.026616
	User Friendly Website	1.481928

Not Applicable	No reason specified	3.969846
	Poor Customer Service	3.988380
	Poor Product Quality	4.009355
	Poor Website	3.966555
	Products always in Stock	1.486842
	Quality Customer Care	1.535977
	Reasonable Price	1.510811
	Too many ads	3.962878
	User Friendly Website	1.502128

Solved	No reason specified	3.997426
	Poor Customer Service	3.964889
	Poor Product Quality	3.998696
	Poor Website	4.042440
	Products always in Stock	1.508982
	Quality Customer Care	1.493750
	Reasonable Price	1.479042
	Too many ads	3.967785
	User Friendly Website	1.583851

Solved in Followup	No reason specified	4.033069
	Poor Customer Service	3.976347
	Poor Product Quality	3.981333
	Poor Website	3.946078
	Products always in Stock	1.542683
	Quality Customer Care	1.503268
	Reasonable Price	1.475177
	Too many ads	3.921543
	User Friendly Website	1.560000

Unsolved	No reason specified	4.027397
	Poor Customer Service	3.966234
	Poor Product Quality	4.001289
	Poor Website	3.913882
	Products always in Stock	1.455090
	Quality Customer Care	1.525714
	Reasonable Price	1.508571
	Too many ads	3.980418
	User Friendly Website	1.506098

Potential Metrics

Precision

Precision judges a classifier based on its ability to not classify a negative label as positive.

Recall

Ratio of TP over (TP+FN). Proportion of actual positives that were identified correctly.

Accuracy

The ratio of correct predictions to all predictions, right or wrong

F1-score

The average that gives equal weight to a set of numbers, reaches its best value at 1 and worst at 0.

AUC

Area under ROC curve, the greater this value the “better” the model. Measures how well model can distinguish between classes.

Why Precision

The cost of mistaking a high risk score for a low one is greater than the reverse. From a business standpoint, a company would rather accidentally classify a loyal customer as one about to leave, rather than not realize that a customer is at a high risk of leaving. This way, businesses have a higher chance of catching customers who are at high risk of leaving, and will also improve the experience of customers that are at a lower risk of leaving.

Using Random Search To Find Best Parameters

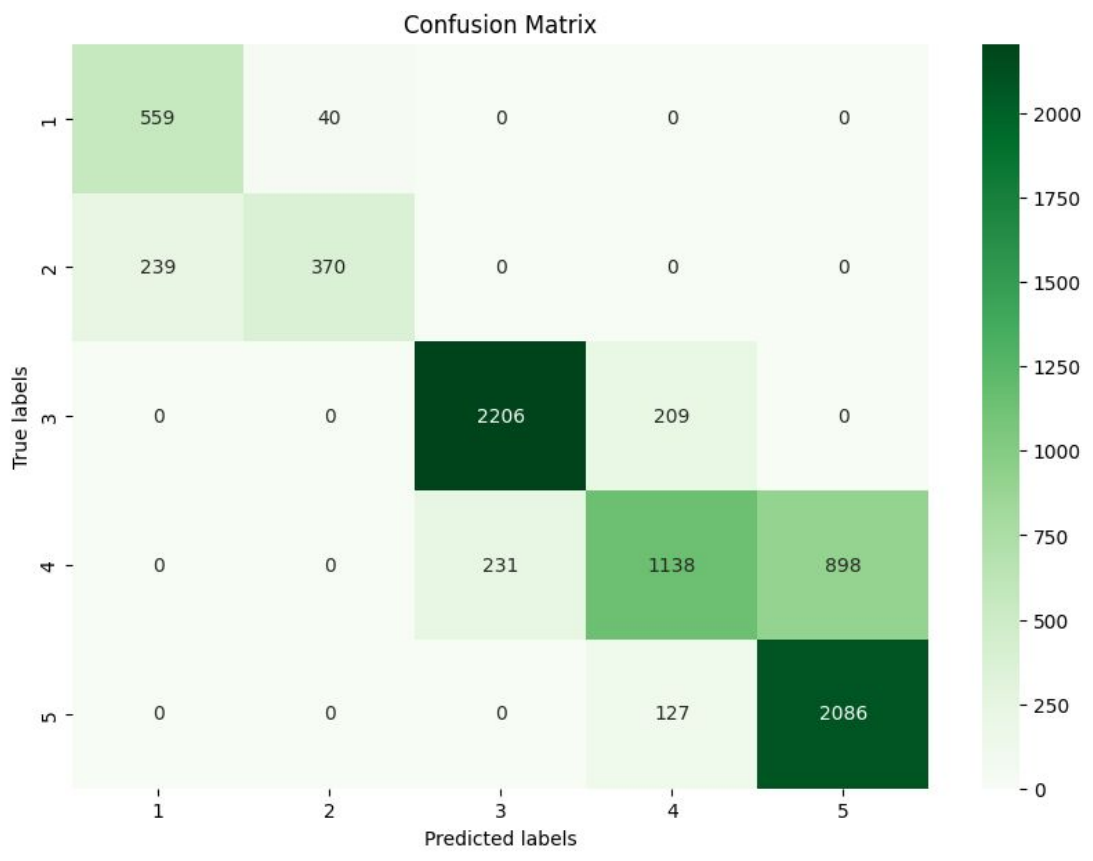
```
from pprint import pprint

max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
random_grid = {'n_estimators' : [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)],
               'max_features' : ['sqrt', 'log2', None],
               'max_depth' : max_depth,
               'min_samples_split' : [2, 5, 10],
               'min_samples_leaf' : [1, 2, 4],
               'bootstrap' : [True, False]}

RF_random = RandomizedSearchCV(estimator = RF_classifier, param_distributions = random_grid,
                               n_iter = 100, cv = 3, verbose = 0, random_state = 42, n_jobs = -1)

RF_random.fit(x_train,y_train)

pprint(RF_random.best_params_)
```

0.3s Python

Best Parameters Found After Running Random Search

```
{'n_estimators': 1600,
'min_samples_split': 2,
'min_samples_leaf': 2,
'max_features': None,
'max_depth': 10,
'bootstrap': True}
```

Confusion Matrix (Random Forest Model)

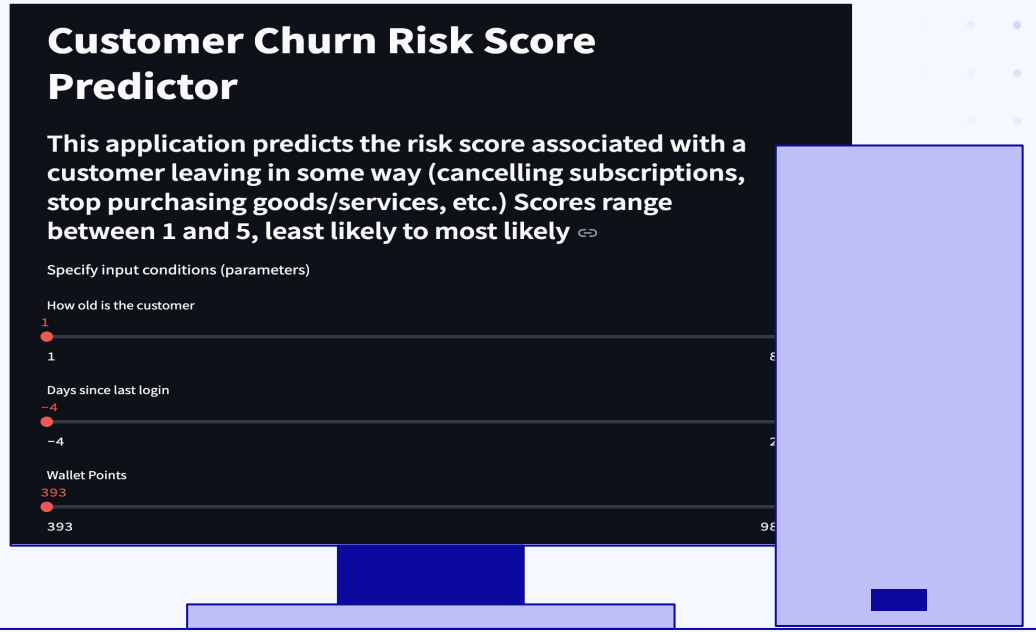


Interpretations

Web App Demo

Making the predictions accessible

https://customer-churn-risk.streamlit.app/



Thanks!

Questions?