Koordinator TSDN 2022:

data academy

CYBERTREND DATA ACADEMY

Asosiasi Data Sains dan AI Indonesia

[Business optimization in predicting customer churn : a machine learning approach]

[ANN Team]

TURNAMEN SAINS DATA NASIONAL 2022

DISPONSORI OLEH :

data academy
CYBERTREND DATA ACADEMY

cybertrend

SAFARINDO

GOOD CHOICE

LSP SAINS DATA DAN KECERDASAN BUATAN INDONESIA

PARAGONCORP

LOJIN Organic Store

karier.mu By sekolah.mu

DIDUKUNG OLEH :

KOMINFO

BADAN PUSAT STATISTIK PROVINSI DKI JAKARTA

BNSP
BADAN NASIONAL SERTIFIKASI PROFESI

BRIN
BADAN RISET DAN INOVASI NASIONAL

JAYA RAYA

JAKARTA open data

JAKARTA SATU
Satu Peta, Satu Data, Satu Kebijakan

JABAR DIGITAL SERVICE

OPEN DATA JABAR

Asosiasi Data Sains dan AI Indonesia

ASOSIASI ILMUWAN DATA INDONESIA

ABDI
ASOSIASI BIG DATA INDONESIA

APTIKOM

DATA SCIENCE INDONESIA

BigData

OUTLINE

ANN TEAM

ANALYTICS : DIGITAL ECONOMI

Nur Amilah | Team

DATA ANALYST

Natalia Dinda S.P | Team

TEACHER

Archie Citra Muhammad | **Team Lead**

FREELANCER DATA ANALYST

TSDN|2022

Archie Citra Muhammad | archiecm09@gmail.com

TTL	: Sragen, 22 Sept 1994
No. Hp	: 08112165945
Address	: Sragen Tengah, Sragen , Jawa Tengah
Social Media	: @archiecm

Nur Amilah|

nuramilahnuramilah@gmail.com

TTL	: Tangerang, 16 May 2001
No. Hp	: 08159887509
Address	: Kp. Pagedangan, Kab. Tangerang, Banten.
Social Media	: @nuramilah_16

Natalia Dinda Sartika Putri | nata.dsptr@gmail.com

TTL	: Tangerang, 09 June 2000
No. Hp	: 085771768020
Address	: Jl. Raya Mauk No.45, Jatiwaringin, Tangerang Regency, Banten.
Social Media	: @nata.dsptr_

FOREWORD

Digital Economy?

(Brynjolfsson & McAfee, 2014)

Business Optimization?

(Apte, 2010)

Machine Learning?

(Al-Sahaf et al., 2019)

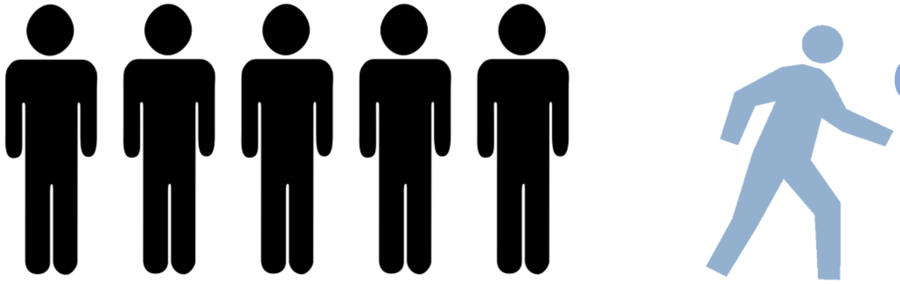
Customer Churn?

(Masarifoglu & Buyuklu, 2019)

Cashback Amount?

(Pinem et al., 2020)

PROBLEM STATEMENT



Churn

16.8%

China Internet Network Information Center (CNNIC)

The E-commerce customer churn rate is up to **80%** compared with traditional business customer management (Wu & Meng, 2016)

Objective

Form a machine learning model with the smallest false negative, identify predictors/factors that influence churn rate and lost opportunity customer churn, and predict customers who have the potential to churn with machine learning models. As well as providing insights & recommendations to identify predictors/factors that influence the churn rate and tenure.

Goals

Predict customer churn rate and provide recommendations to the business team so the company can implement a customer retention strategy.

Business Metrics

$$\text{Churn Rate} = \frac{\text{CUSTOMER CHURN}}{\text{TOTAL CUSTOMERS}}$$

Lost Opportunity

Exploratory Data Analysis (EDA)

Using **Correlation Matrix, Bivariate and Multivariate**

DATA OVERVIEW

```
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   CustomerID                   5630 non-null   int64
 1   Churn                        5630 non-null   int64
 2   Tenure                       5366 non-null   float64
 3   PreferredLoginDevice         5630 non-null   object
 4   CityTier                     5630 non-null   int64
 5   WarehouseToHome              5379 non-null   float64
 6   PreferredPaymentMode         5630 non-null   object
 7   Gender                       5630 non-null   object
 8   HourSpendOnApp               5375 non-null   float64
 9   NumberOfDeviceRegistered     5630 non-null   int64
 10  PreferedOrderCat             5630 non-null   object
 11  SatisfactionScore            5630 non-null   int64
 12  MaritalStatus                5630 non-null   object
 13  NumberOfAddress              5630 non-null   int64
 14  Complain                     5630 non-null   int64
 15  OrderAmountHikeFromlastYear  5365 non-null   float64
 16  CouponUsed                   5374 non-null   float64
 17  OrderCount                   5372 non-null   float64
 18  DaySinceLastOrder            5323 non-null   float64
 19  CashbackAmount               5630 non-null   float64
dtypes: float64(8), int64(7), object(5)
```

Source : https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

Target Variable :

Churn (Classification Model)

Tenure (Regression Model)

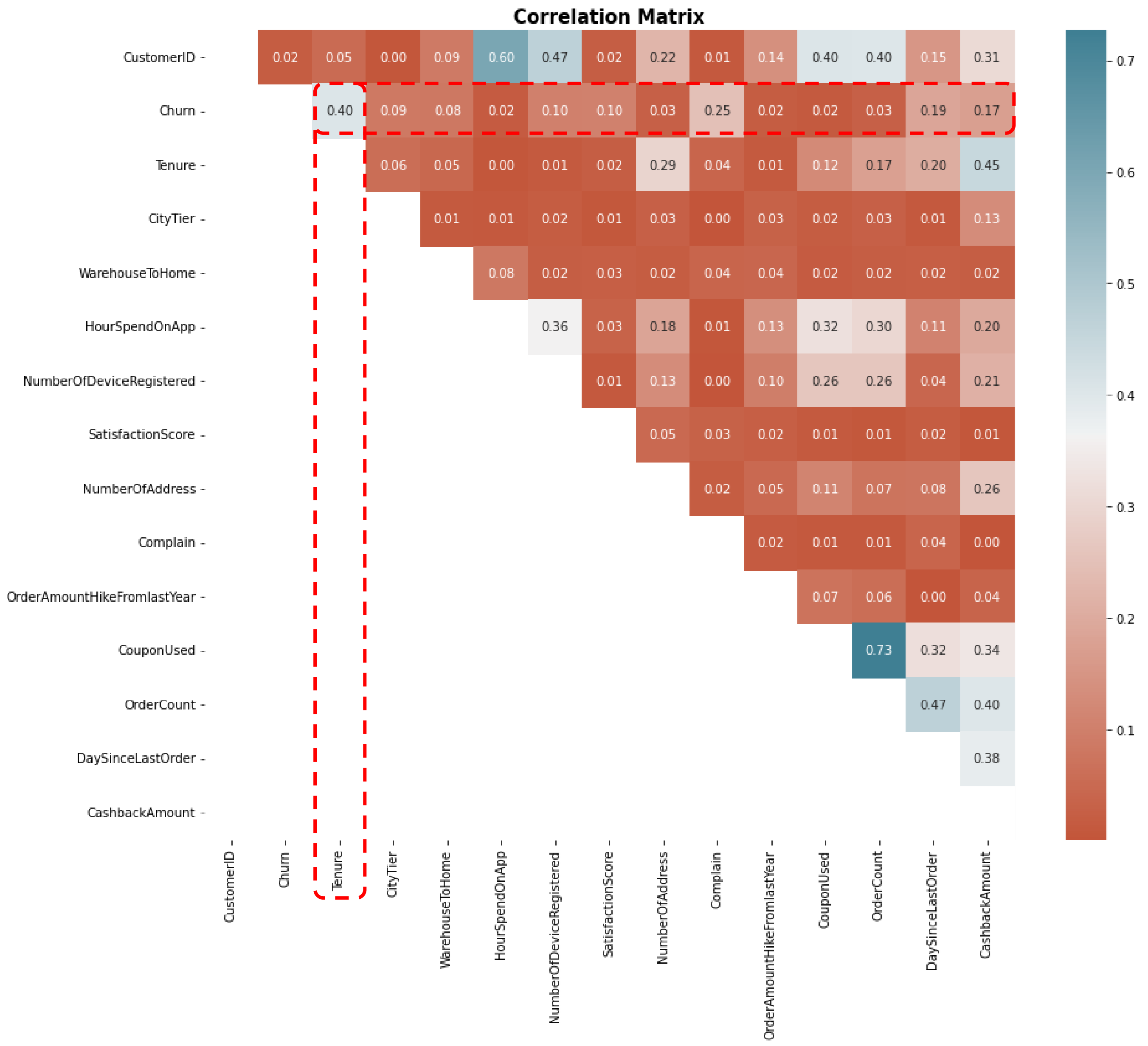
Dataset Information?

(20 kolom dan 5630 baris, 19 variabel input, 2 var.target)

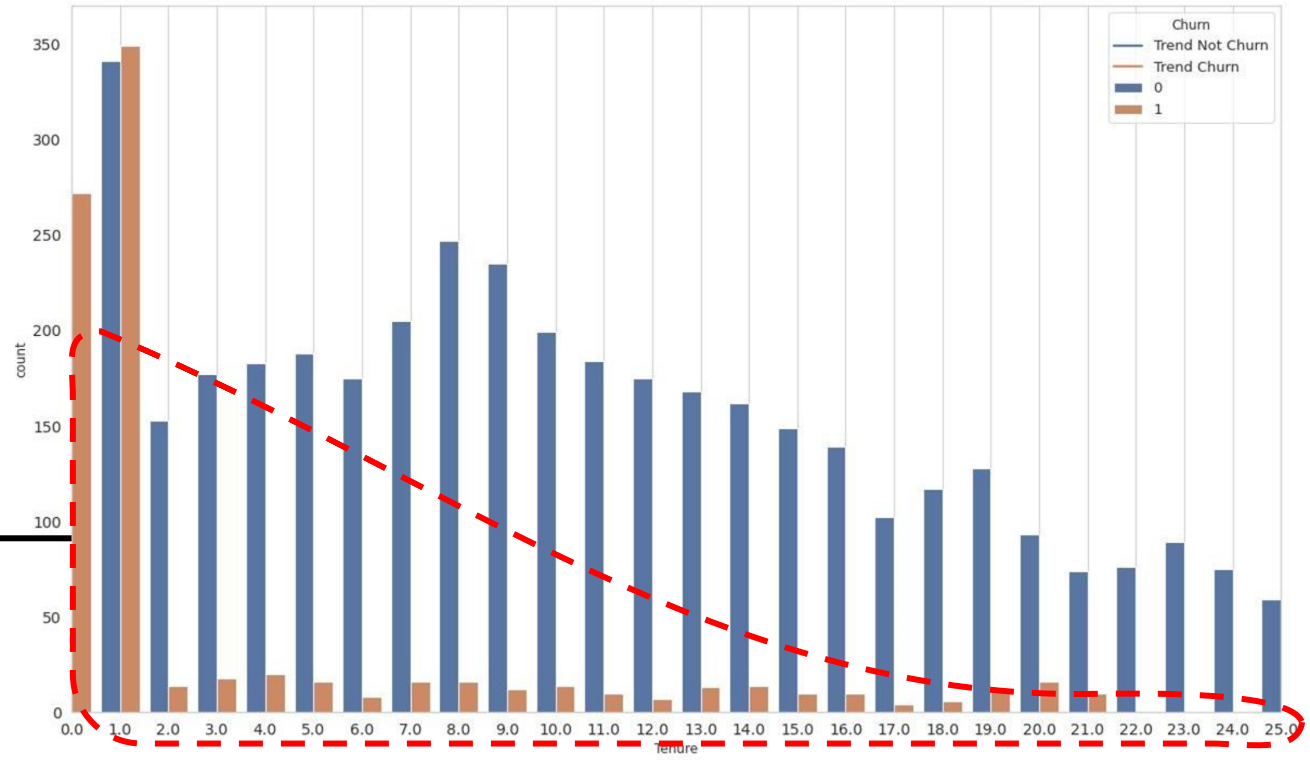
Dtype (Data Type)

EXPLORATORY DATA ANALYSIS

Column	Correlation_ratio
Tenure	0.40
Complain	0.25
Cashbackamount	0.17
Daysincelastorder	0.19
Numberofdeviceregistered	0.10
Satisfactionscore	0.10
Citytier	0.09
Warehousetohome	0.08
Numberofaddress	0.03
Ordercount	0.03
Hourspendonapp	0.02
Couponused	0.02
Orderamounthikefromlastyear	0.02

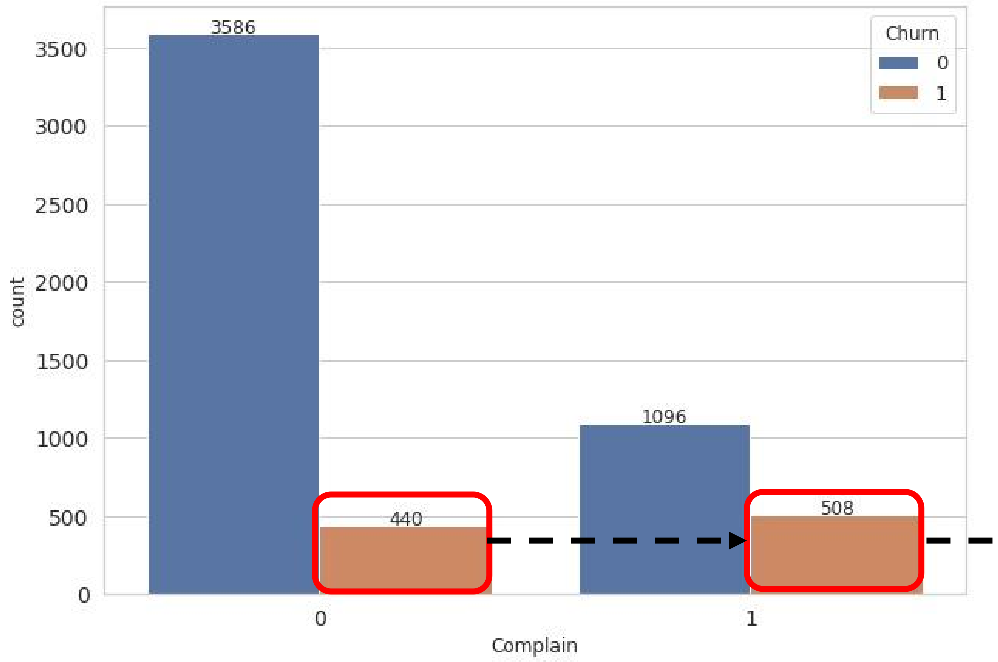


INSIGHTS
(Churn and Not Churn)



The longer tenure, the lower number of churns. And not churn has a steeper trend compared to Churn.

INSIGHTS
(Comparison Complain to Churn and Not Churn)



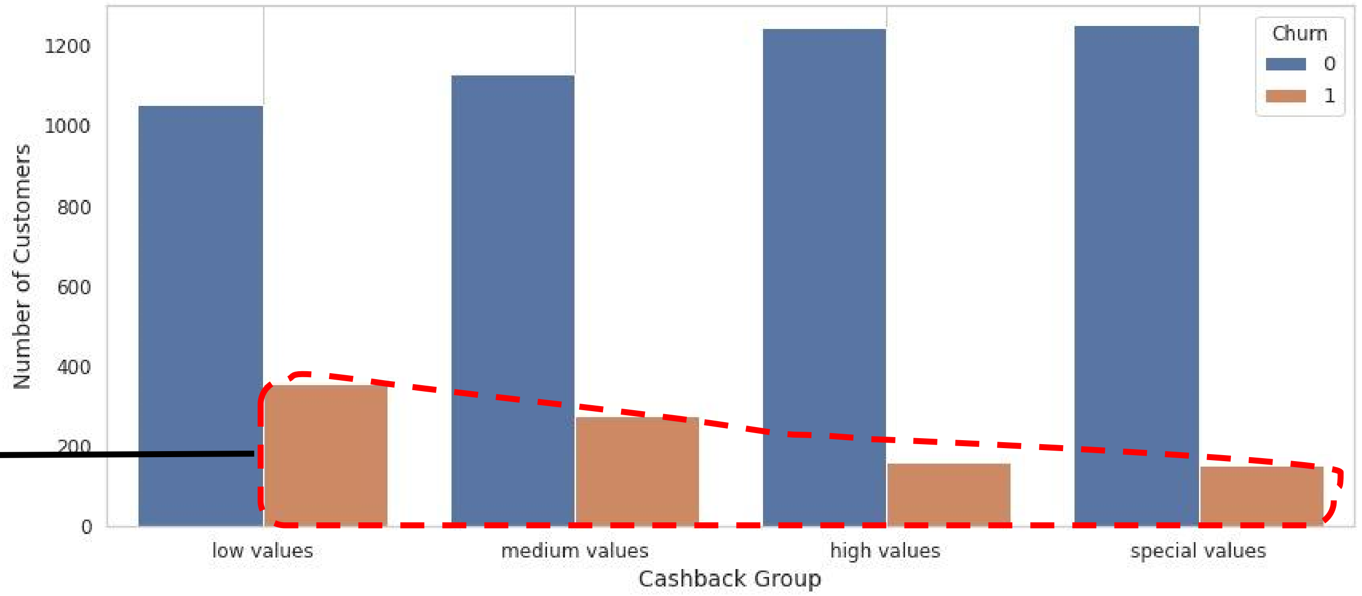
1. Customers with the **highest churn** of **9.0%** are on **customer complaints**.

2. Customers with the **lowest churn** of **7.8%** are **non-complaining customers**.

The more customer complaints increase, the higher the churn rate.

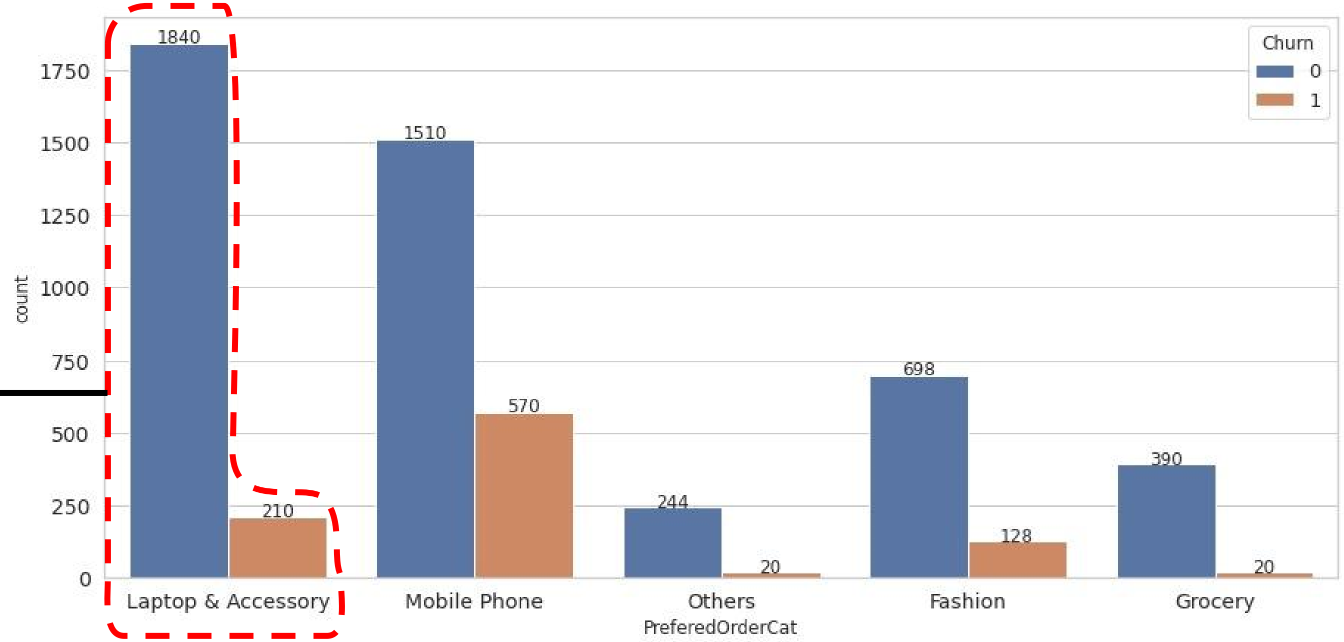
INSIGHTS
(Distribution of Cashback Customers)

Increase Cashback Amount has trend Positive in Not Churn On the contrary Increase Cashback Amount has trend Negative in Churn



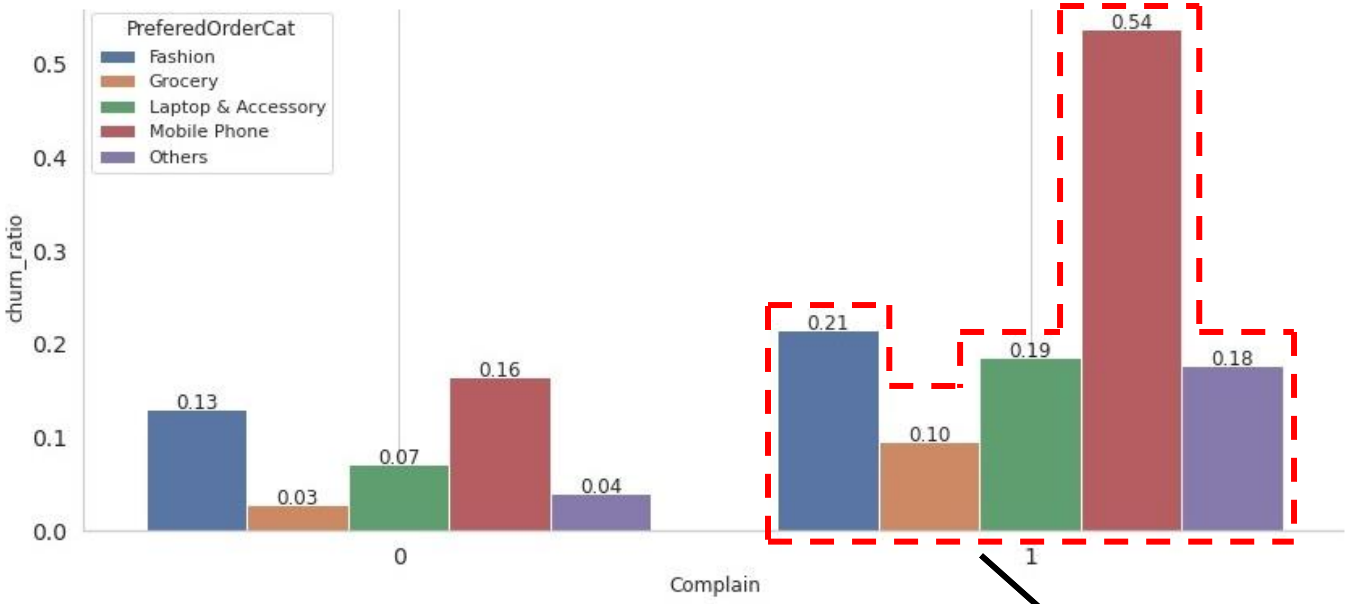
INSIGHTS
(Prefered Order Categories Customer)

Customer who ordered Laptop and Accessory has a significant number of Not Churn compared same order category with Churn.



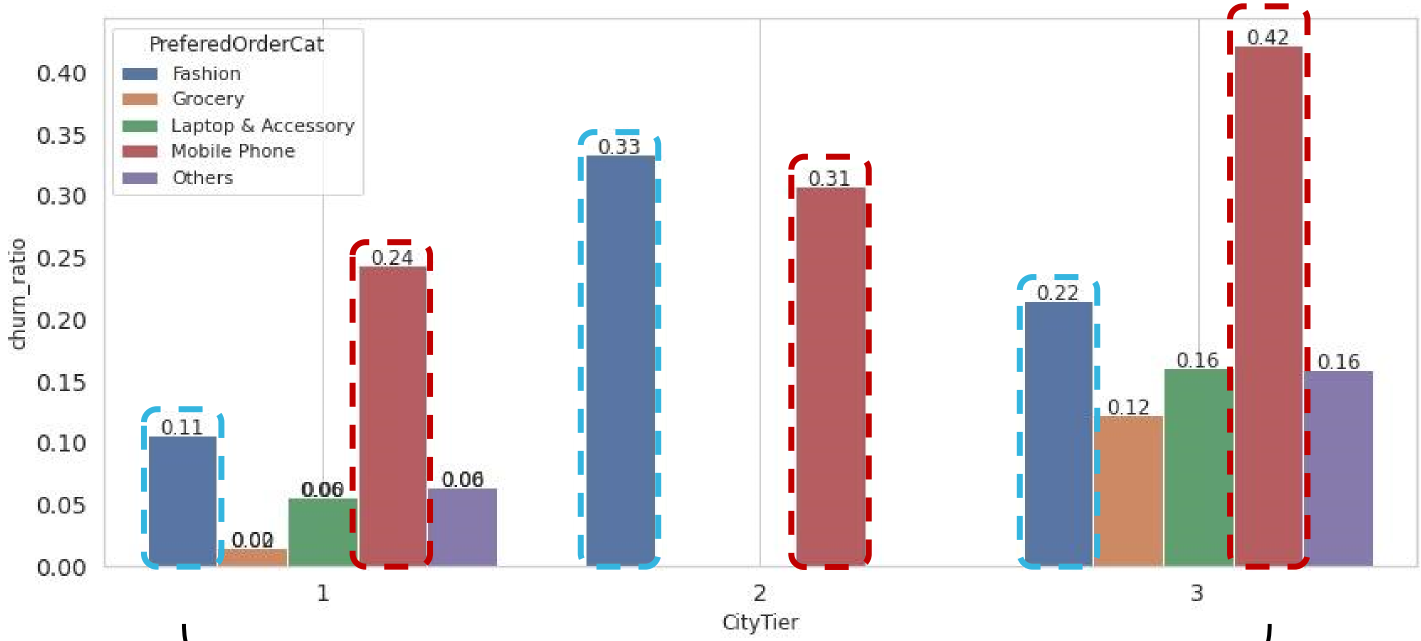
INSIGHTS

(Distribution of Complain & Order Categories vs Ratio Churn)



Customers with complaints have a ratio churn increase in all order categories.

INSIGHTS
(Distribution of Complain & Order Categories vs Ratio Churn)

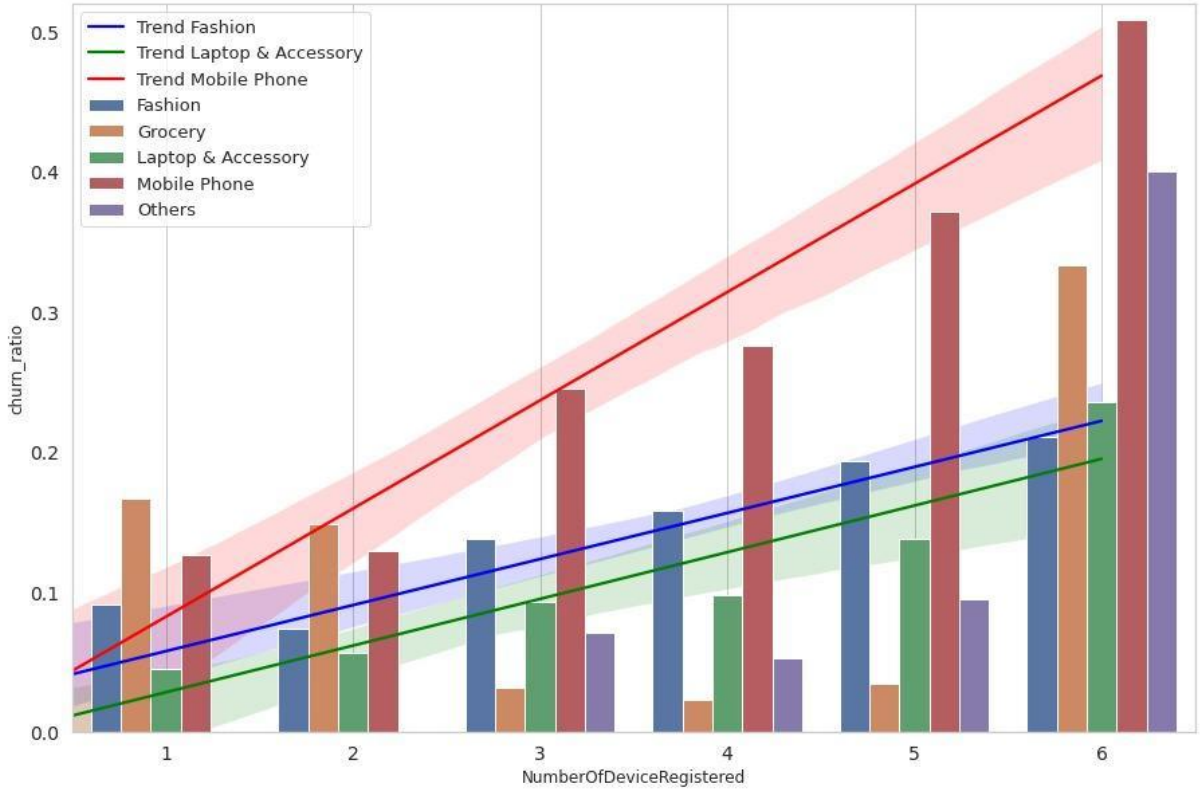


More City Tier increase, more ratio churn increase in Fashion and Mobile Phones

INSIGHTS
(Distribution of Complain & Order Categories vs Ratio Churn)

“

More Number Of Device Registered increased and more ratio churn increased in Fashion, laptops & accessories, and Mobile Phones.



INSIGHTS
(External Data)

85% of customer churn because of **poor service** that could have been prevented.

Source : https://www.slideshare.net/ekolsky/cx-for-executives

82% of customers have stopped doing business with e-commerce because of **bad customer service**.

Source : https://www.zendesk.com/blog/whycompanies-should-invest-in-the-customer-experience/

67% of customer churn could be avoided if **resolved the customer's issue** during their first interaction.

Source : https://www.getfeedback.com/resources/cx/40-stats-churn-customersatisfaction/#:~:text=67%25%20of%20customer%20churn%20could,(Kolsky)

Data Pre-Processing

Data Cleaning

Check Irrelevant Data
Check Missing Data
Check Duplicate
Check Outlier

Feature Encoding

One Hot Encoder

Simple Imputer

Iterative Imputer

Transforming

80 : 20
Train : Test

Feature Encoding

Robust Scaler

Standard Scaler

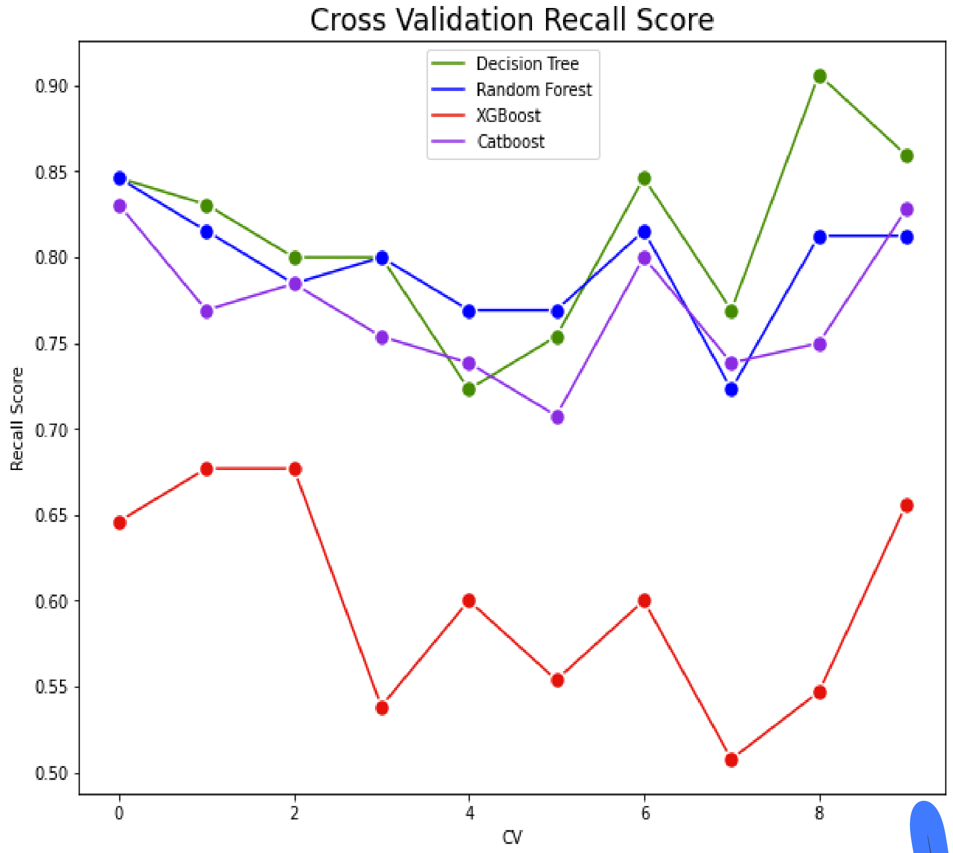
Predict Churn

Selection Models & Cross Validation, Handling Imbalance, Hyperparameter Tuning, Feature Importance with SHAP

Model Selection & Cross-Validation

Models	Recall	Cross Validation	Standar Deviasi
Decision Tree	0.863095	0.803951	0.038707
Catboost	0.809524	0.780246	0.047106
Random Forest	0.797619	0.760799	0.048173
Xgboost	0.553571	0.609570	0.066454
Logistic Regression	0.476190	0.530575	0.051384

NB : Due to an imbalance dataset



Handling Imbalance Target

NB : Due to imbalance dataset

CatBoost

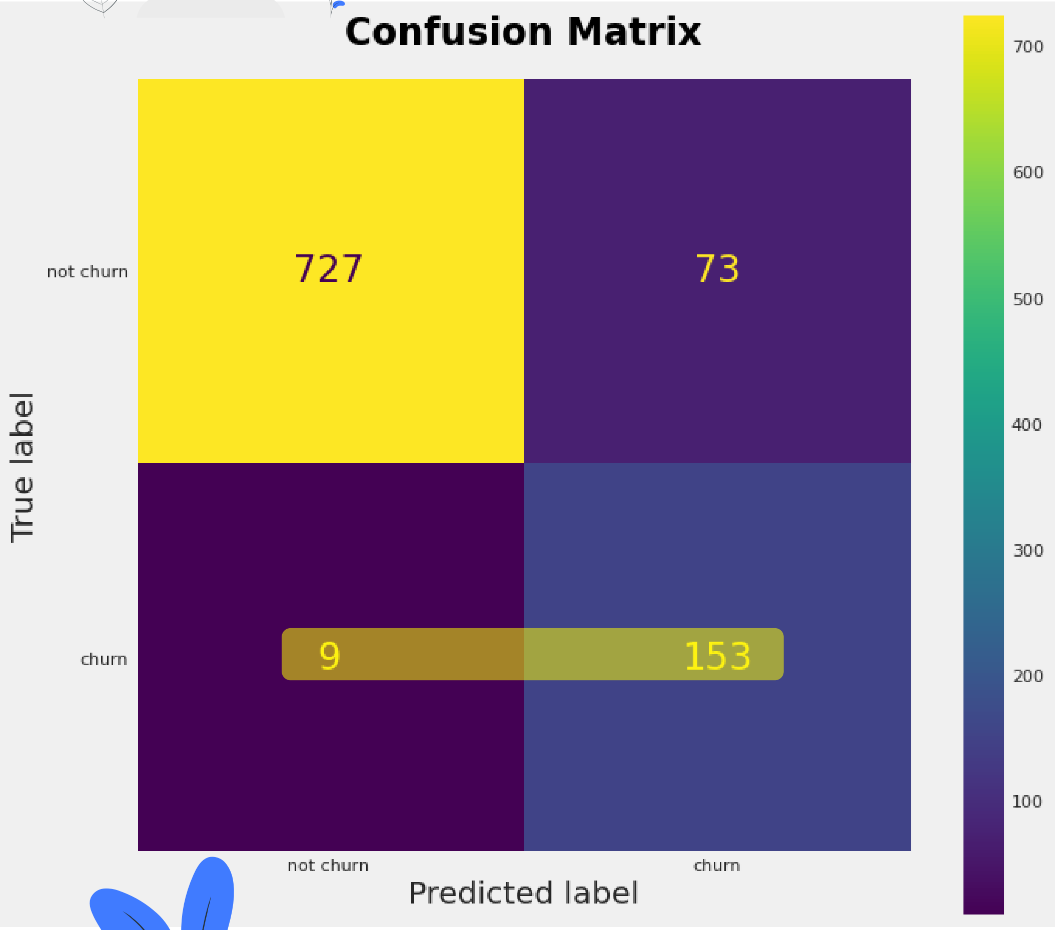
	Without	Undersampling	Oversampling
Train Recall	0.953360	0.996913	0.999826
Test Recall	0.784038	0.928990	0.915144

DECISION TREE

	Without	Undersampling	Oversampling
Train Recall	1.000000	1.000000	1.000000
Test Recall	0.836538	0.881202	0.819519

Catboost have **best fit** in **undersampling and oversampling**. But we choose **undersampling** because it has **gap (train-test) smaller than other**.

CatBoost Classifier + Undersampling



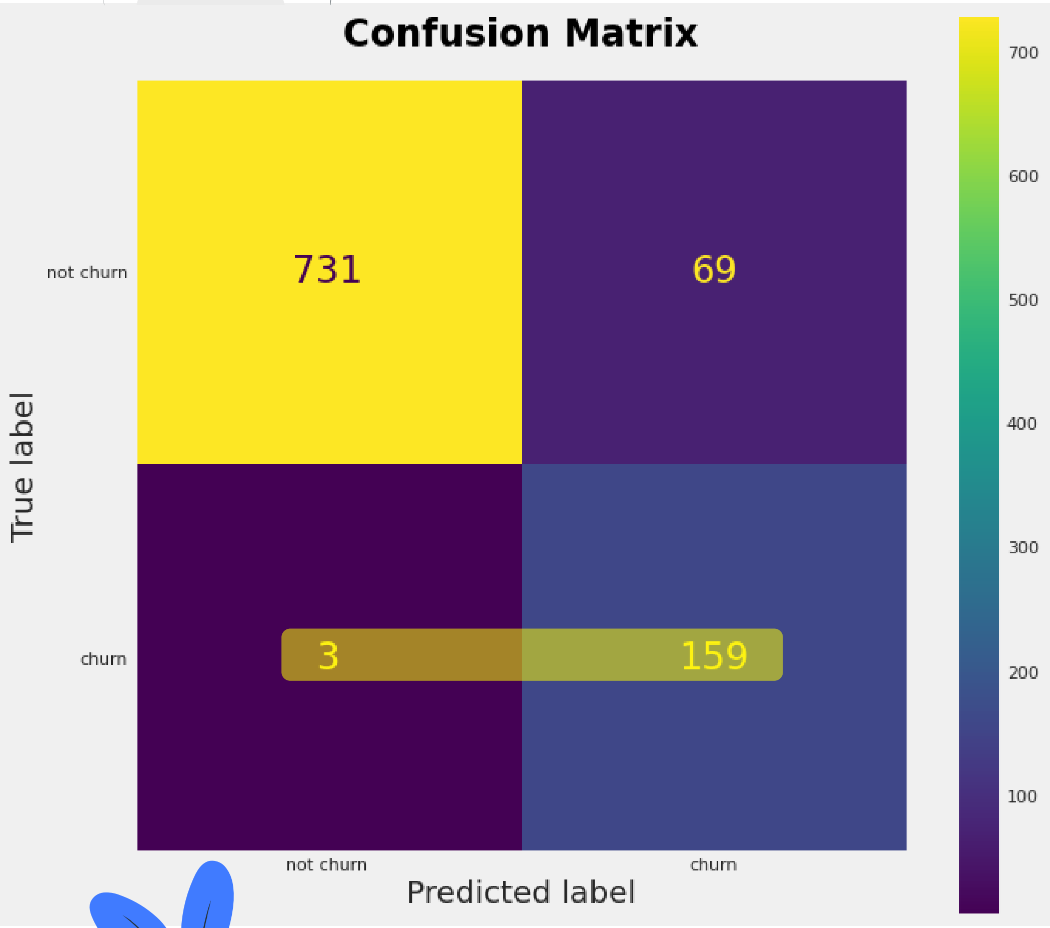
classification_report before tuning:

	precision	recall	f1-score	support
0	0.99	0.91	0.95	800
1	0.68	0.94	0.79	162
accuracy			0.91	962
macro avg	0.83	0.93	0.87	962
weighted avg	0.94	0.91	0.92	962

Recall 0.9444444444444444

Recall ; How many customers did we **correctly predict** to take an interest with our product compared to all customers which are **truly churn**? **94%**

CatBoost Classifier + Undersampling + Tuning



n_estimators = [1000]

learning_rate = [float(x) for x in np.linspace(0.001, 0.1, 20)]

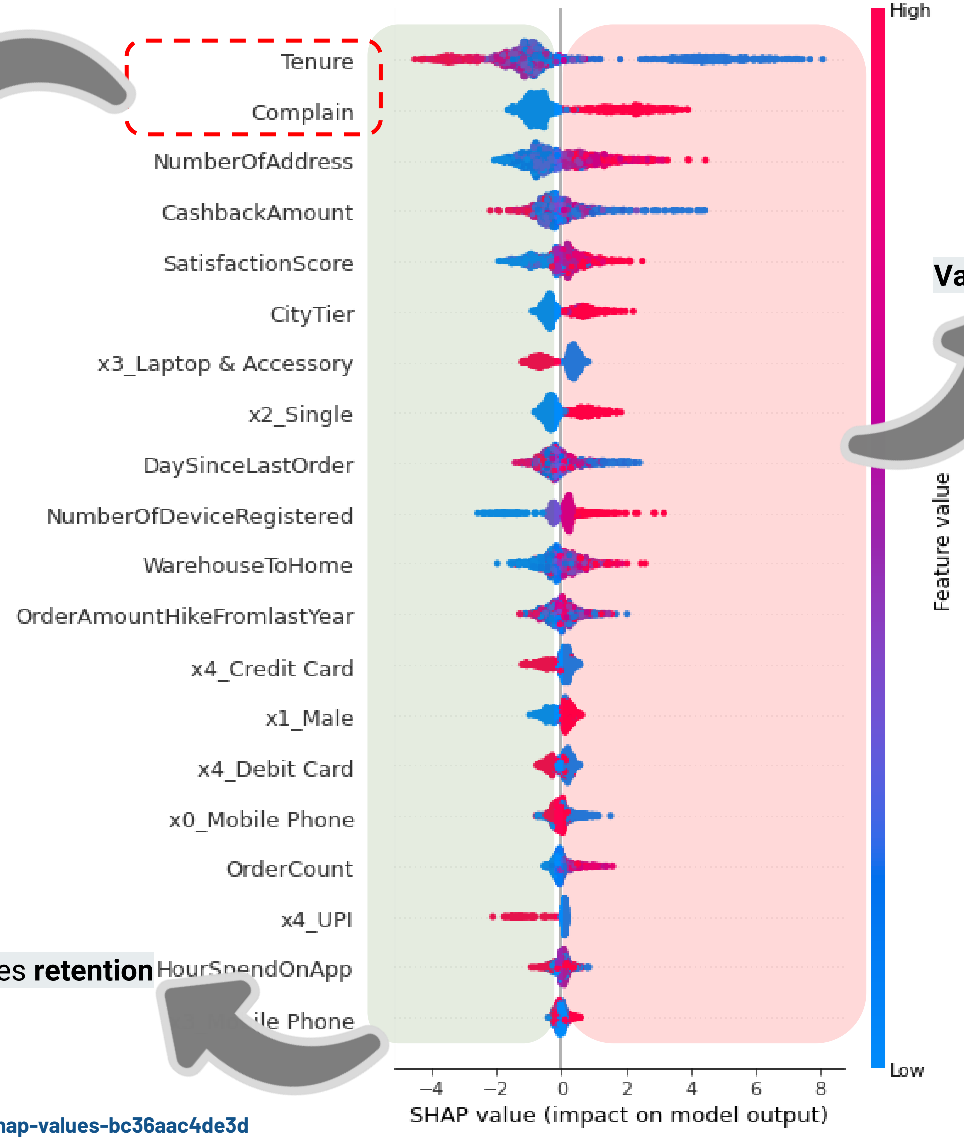
```
classification_report after tuning:
              precision    recall  f1-score   support

           0       1.00      0.91      0.95       800
           1       0.70      0.98      0.82       162

    accuracy                           0.93       962
   macro avg       0.85      0.95      0.88       962
weighted avg       0.95      0.93      0.93       962
```

Recall 0.9814814814814815

Recall 98%

Feature Importance with SHAP

Feature Importance will be exploratory



Value that generates **churn**

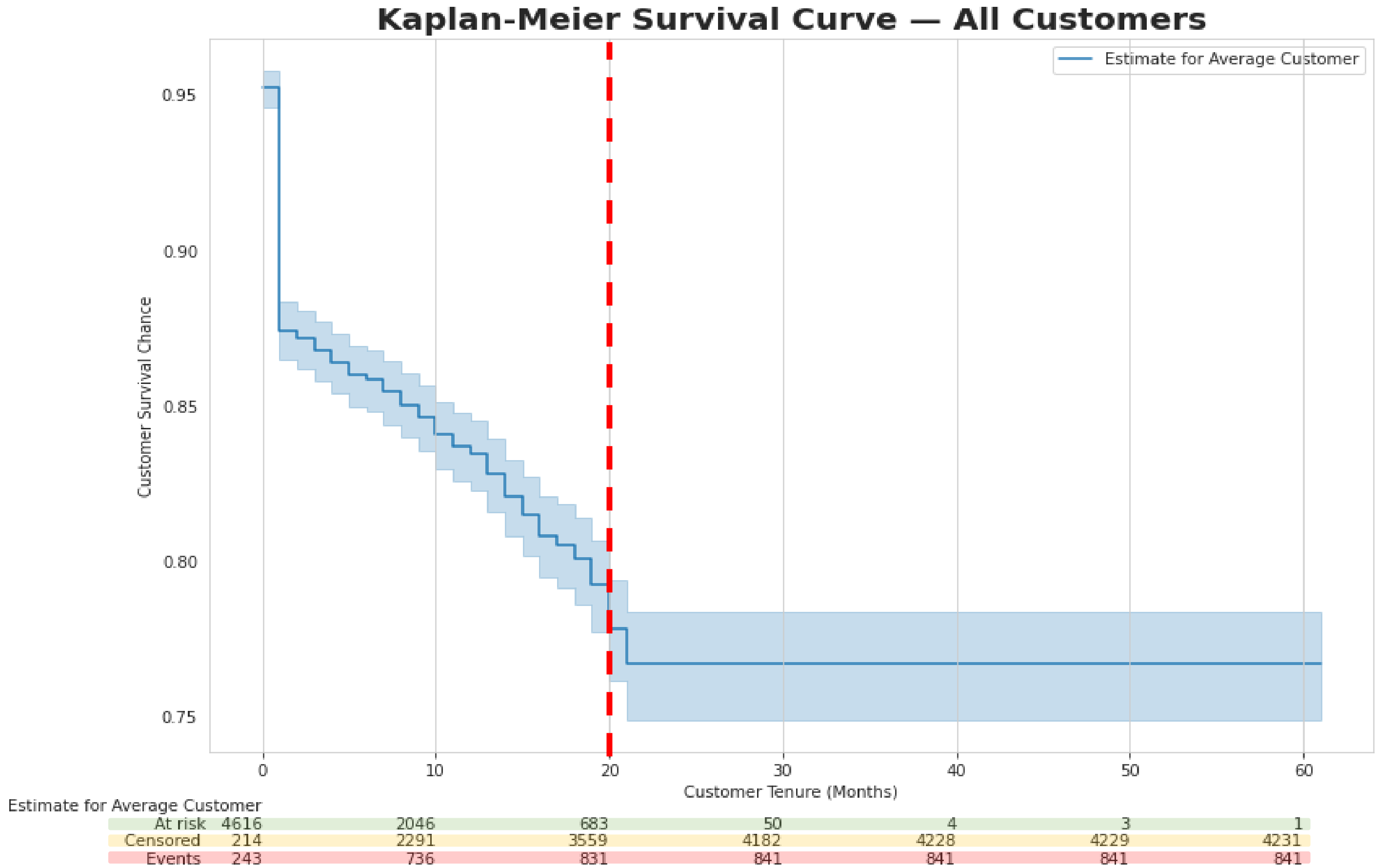
Value that generates **retention**

https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d

Survival Analyst

Using Kaplan-Meier (KM) and COx Proportional Hazard (CPH) Model

Kaplan-Meier(KM) Survival Curve



20 → **At risk : 683** customer **not churn (13%)**

20 → **Censored : 3559** customer **will be churn (70%)**

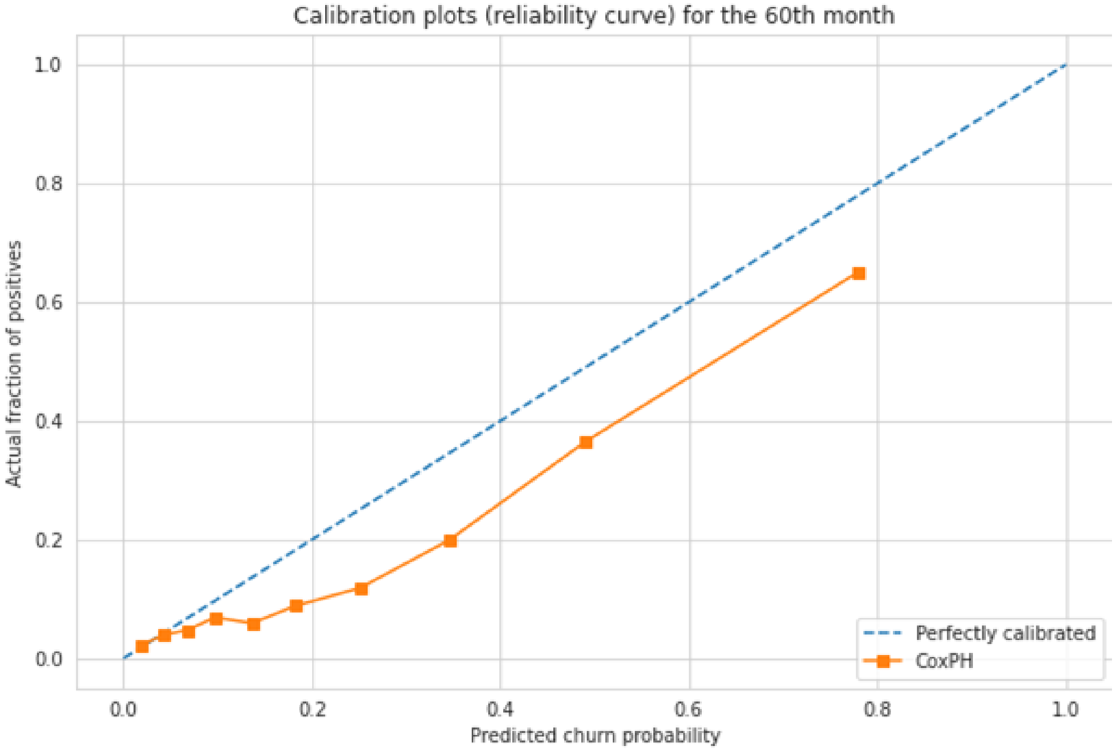
20 → **Events : 831** customer **churn (17%)**

https://towardsdatascience.com/how-to-not-predict-and-prevent-customer-churn-1097c0a1ef3b

Cox Proportional Hazard (CPH) Model

Model	lifelines.CoxPHFitter
Duration col	'Tenure'
Event col	'Churn'
Baseline estimation	breslow
Number of observations	5073
Number of events observed	841
Partial log-likelihood	-6296.226
Time fit was run	2022-11-19 08:47:34 UTC
Model	base model
Concordance	**0.829**
Partial AIC	12640.452
log-likelihood ratio test	1223.310 on 24 df
-log2(p) of ll-ratio test	805.834

AUC-ROC logistic regression.

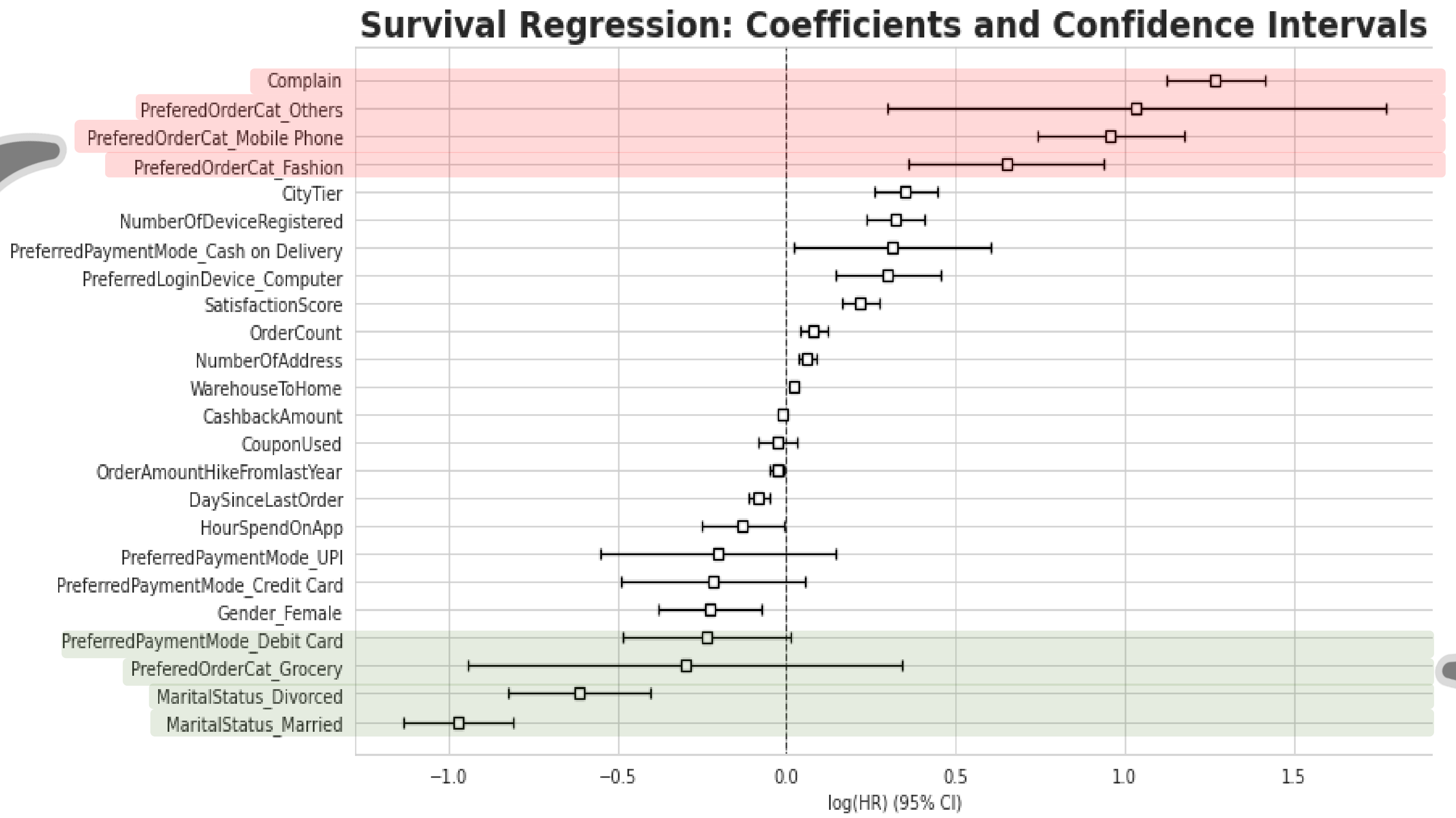


The Brier Score of our **CPH Model is 0.11** at the end of **60 months**

Predictions from the **model up to 60 months** are still **close to true values.**

https://towardsdatascience.com/how-to-not-predict-and-prevent-customer-churn-1097c0a1ef3b

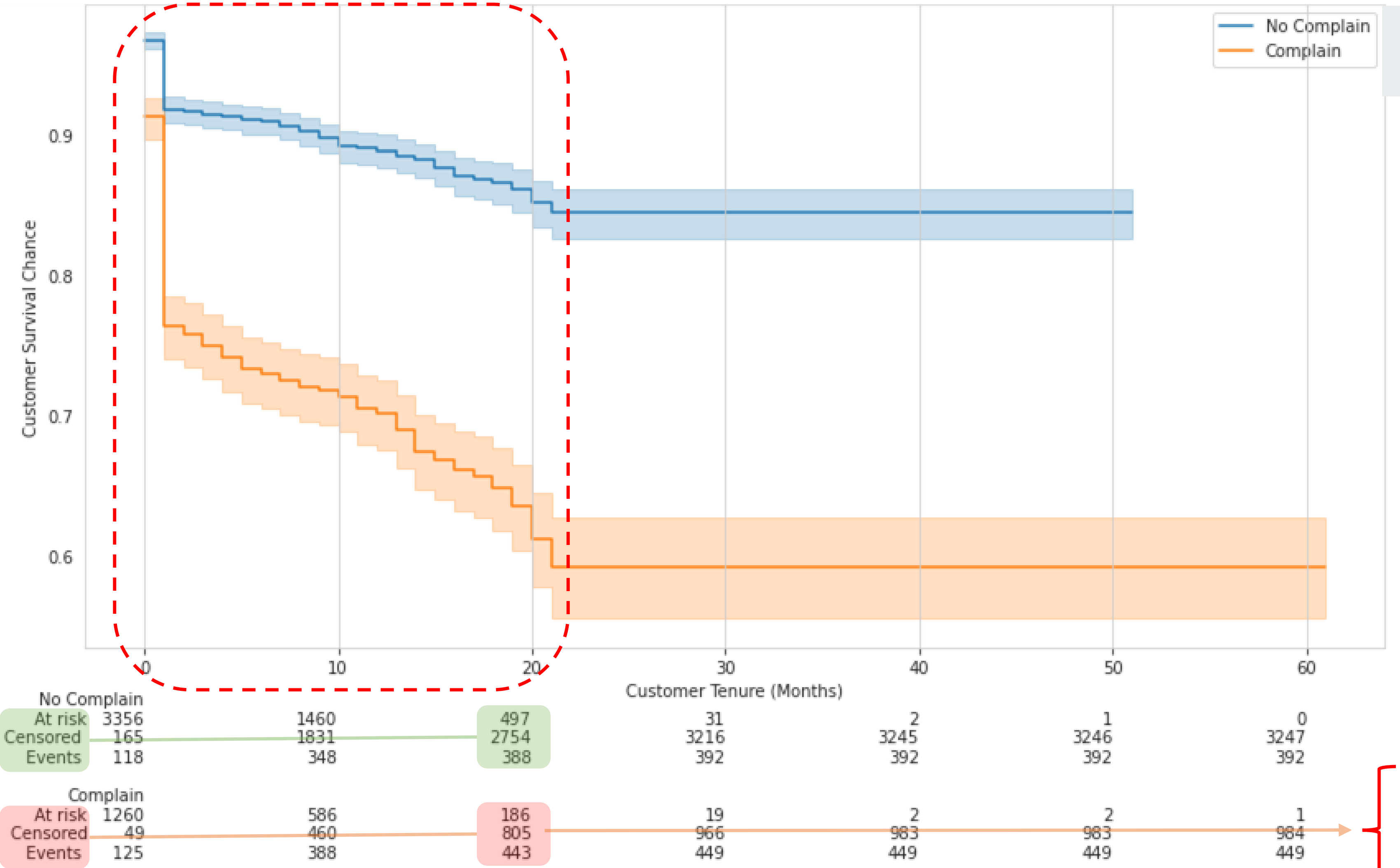
CPH Model Visualization



Feature that generates **churn**

Feature that generates **retention**

Churn Prediction and Prevention



15%
75%
10%

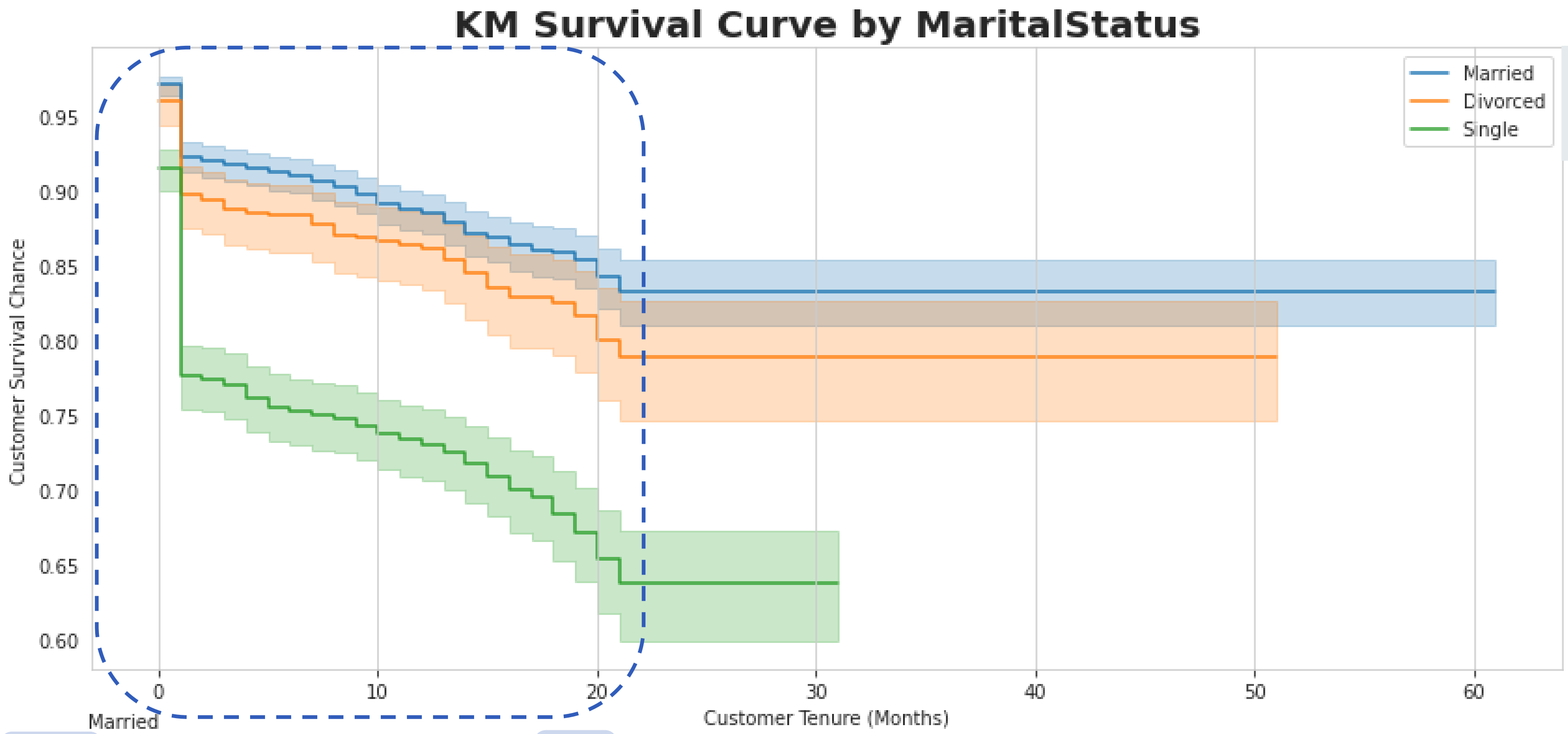
15%
56%
29%

Customer Survival Chance

No complaint have **89%**

Complaint have **68%**

Churn Prediction and Prevention



		0	10	20	30	40	50	60
Married	At risk	2494	1161	377	29	2	2	1
	Censored	101	1254	1997	2341	2368	2368	2369
	Events	77	257	298	302	302	302	302
Divorced	At risk	778	368	141	10	2	1	0
	Censored	36	377	585	714	722	723	724
	Events	34	103	122	124	124	124	124
Single	At risk	1344	517	165	11	0	0	0
	Censored	77	660	977	1127	1138	1138	1138
	Events	132	376	411	415	415	415	415

15%
75%
14%

11%
63%
26%

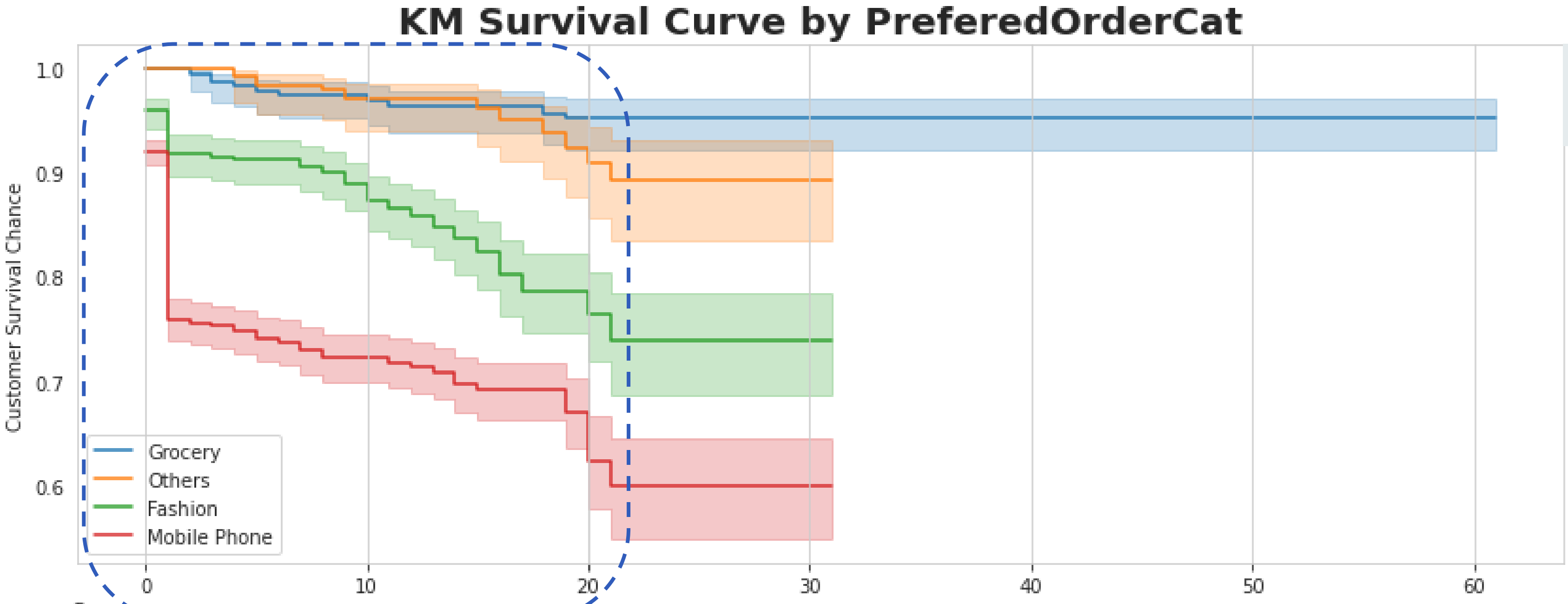
Customer Survival chance in Marital Status

88% of Married

85% of Divorced

73% of Single

Churn Prediction and Prevention



Grocery							
	At risk	365	339	199	11	2	2	1
	Censored	1	16	151	339	348	348	349
	Events	0	11	16	16	16	16	16
Others							
	At risk	241	233	111	5	0	0	0
	Censored	1	2	114	218	223	223	223
	Events	0	7	17	19	19	19	19
Fashion							
	At risk	724	397	120	12	0	0	0
	Censored	8	278	523	627	639	639	639
	Events	31	88	120	124	124	124	124
Mobile Phone							
	At risk	1582	375	101	9	0	0	0
	Censored	125	1020	1270	1358	1367	1367	1367
	Events	148	460	484	488	488	488	488

54%
41%
5%

6%
68%
26%

Customer Survival chance in Prefered Order Category

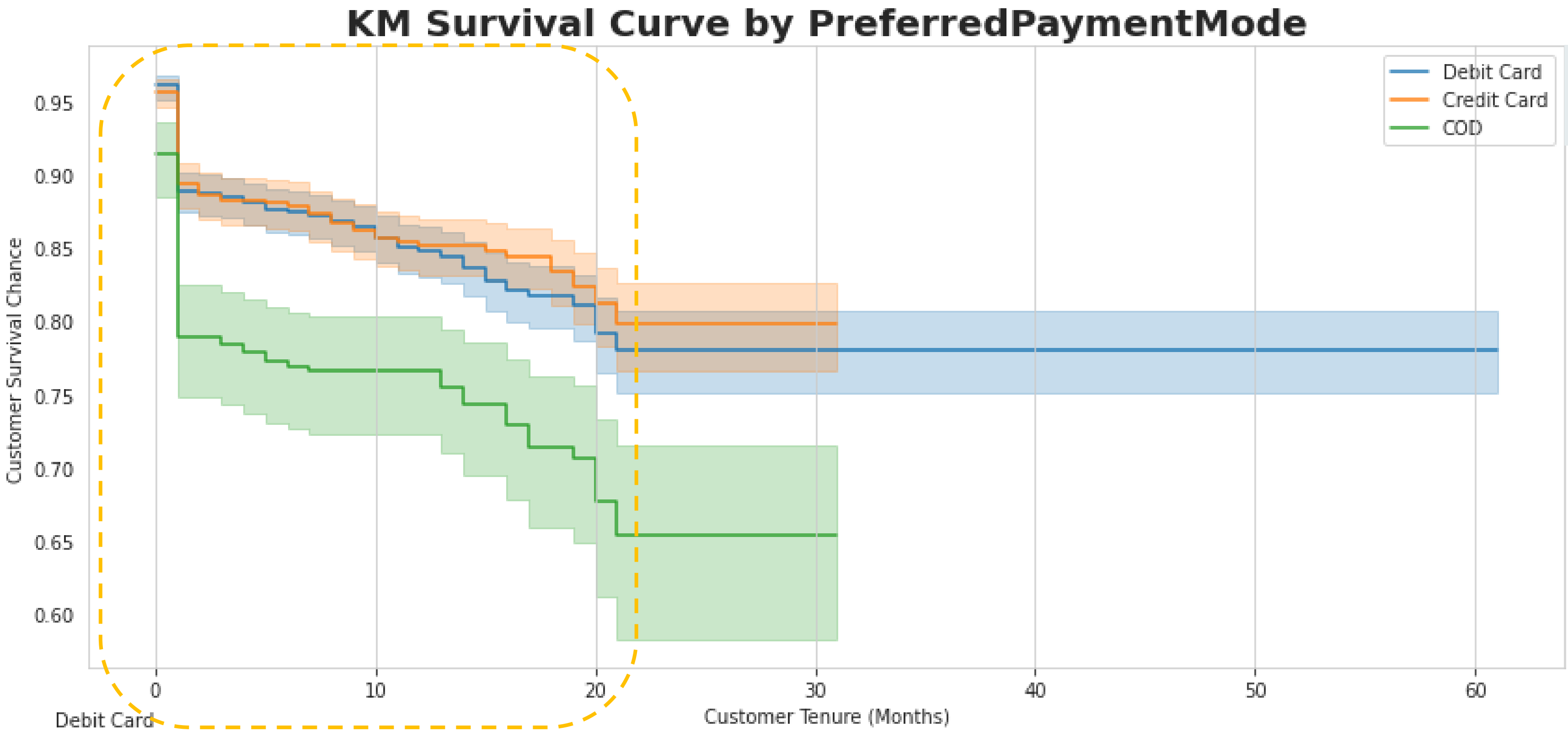
95% of Grocery

92% of Others

83% of Fashion

73% of Mobile Phone

Churn Prediction and Prevention



	0	10	20	30	40	50	60
Debit Card							
At risk	1932	860	273	22	4	3	1
Censored	85	967	1514	1761	1779	1780	1782
Events	81	271	311	315	315	315	315
Credit Card							
At risk	1450	647	240	16	0	0	0
Censored	78	747	1134	1354	1370	1370	1370
Events	68	202	222	226	226	226	226
COD							
At risk	402	162	59	3	0	0	0
Censored	16	193	284	338	341	341	341
Events	39	102	114	116	116	116	116

15%
71%
14%

13%
62%
25%

Customer Survival chance in Prefered Payment Mode

84% of Debit Card

86% of Credit Card

74% of COD

Recommendation Business

1

Complain

Create **Customer Experience**, can **Fast Handling** Complain, ask for customer **feedback at regular intervals**

2

Marital Status

Targeting campaigns, products and services fit with status Married or **Family Package**

Preferred Payment Mode

Encourage customers to set up payments either through a **debit card or credit card**

3

4

Preferred Order Cat

Encourage customers to preferred **order category grocery**

https://www.zendesk.com/blog/why-companies-should-invest-in-the-customer-experience/

predict_survival_function

Censored Customer with Tenure 60 months

https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html#lifelines.fitters.coxph_fitter.SemiParametricPHFitter.predict_survival_function

Calculate Expected Loss Table

CustomerID	Cashback Amount	Exp_Churn_Month	Exp_Loss	Baseline
50046	130.58	11.00	1,436.38	11.00
50048	120.88	19.00	2,296.72	19.00
50177	112.00	15.00	1,680.00	15.00
50194	124.78	14.00	1,746.92	14.00
50230	147.36	14.00	2,063.04	14.00

CustomerID 50046 :

Churn at month **11**

Expected Loss of **$14,363**

Estimated Revenue Uplift Table

OrderCat_Grocery_Uplift	PaymentMode_Credit Card_Uplift	PaymentMode_Debit Card_Uplift
16.00	14.00	15.00
20.00	19.00	20.00
20.00	15.00	20.00
19.00	17.00	14.00
17.00	16.00	17.00

Estimated **Revenue Uplift**

Order Category **Grocery $ 160**

Payment **Debit Card $ 150**

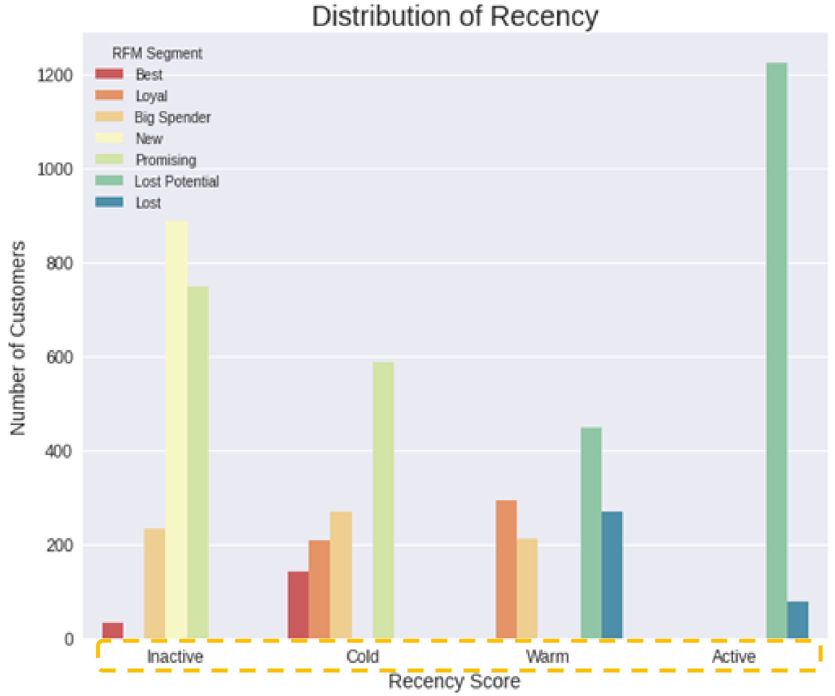
Payment **Credit Card $ 140**

Segmentation of Customer

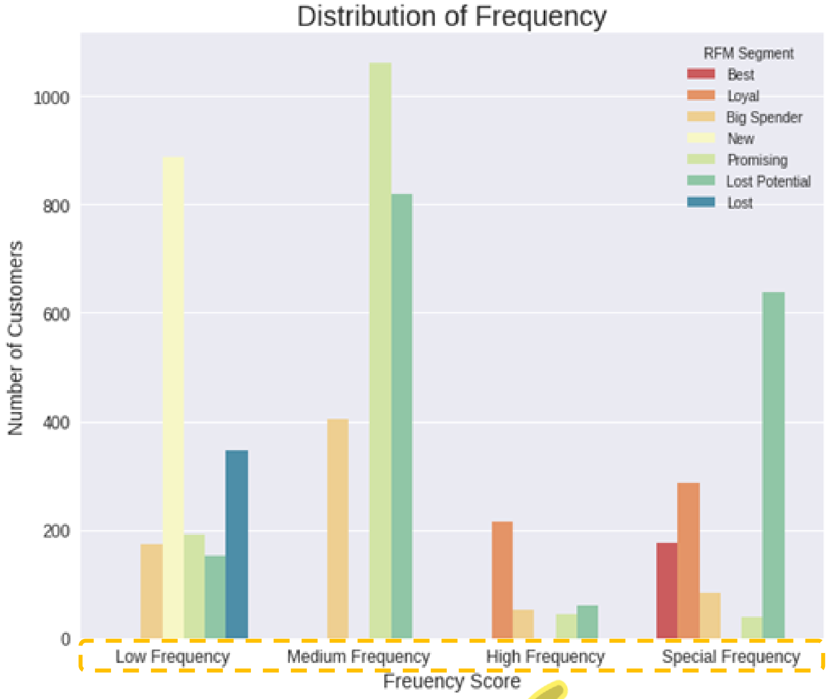
Using **RFM Segmentation**, **K-Means**, and **Gaussian**

RFM Segmentation

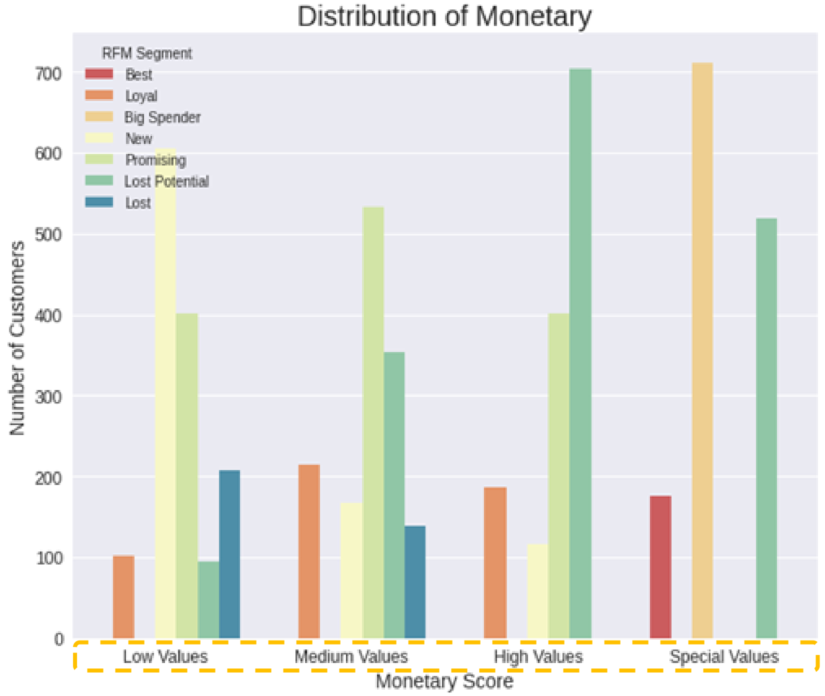
Column **DaySinceLastOrder** as **Recency**



Column **OrderCount** as **Frequency**

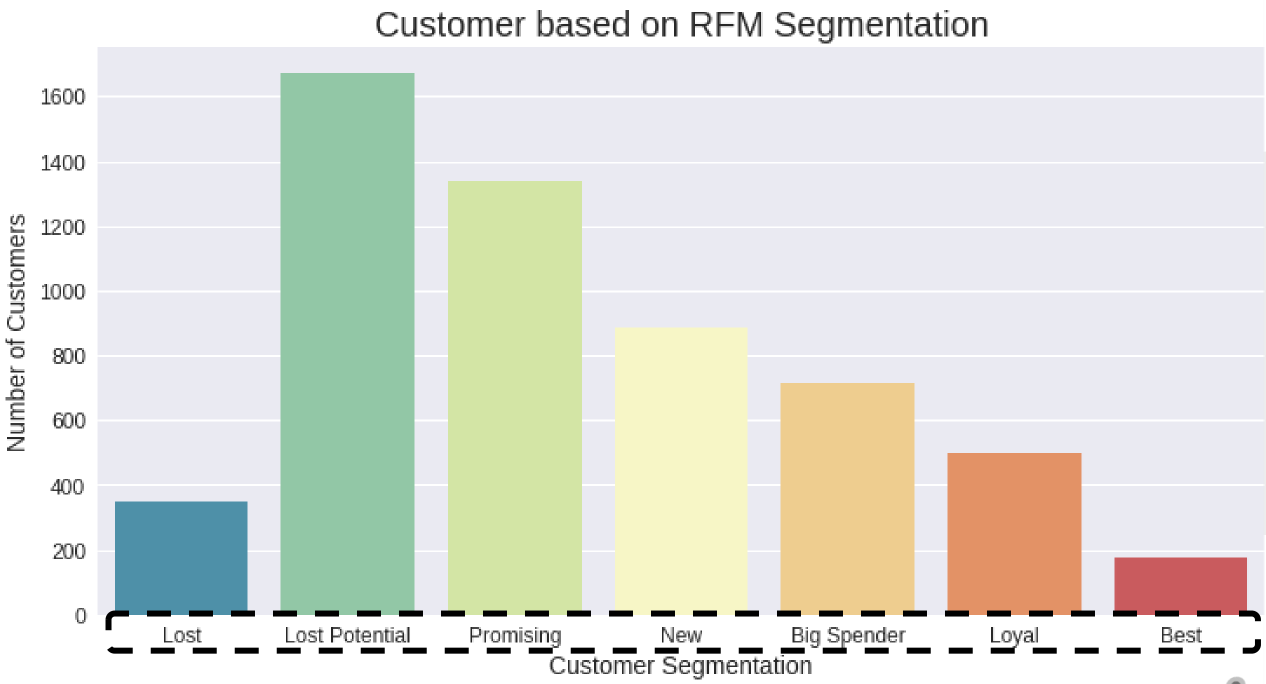


Column **CashbackAmount** as **Monetary**



Segment RFM Score

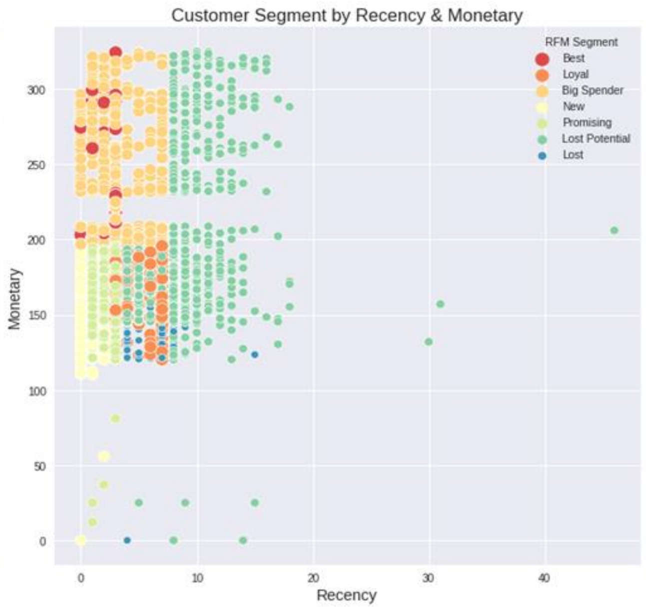
RFM Segmentation

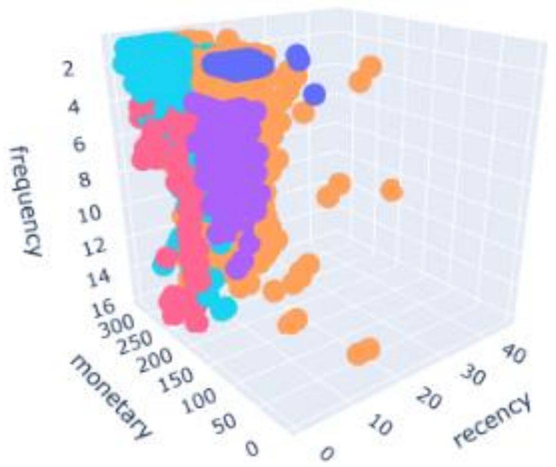


Best Loyal Big Spender New Promising Lost Potential Lost

The **RFM segment** is **based on** scores from the **Recency, Frequency,** and **Monetary distribution**

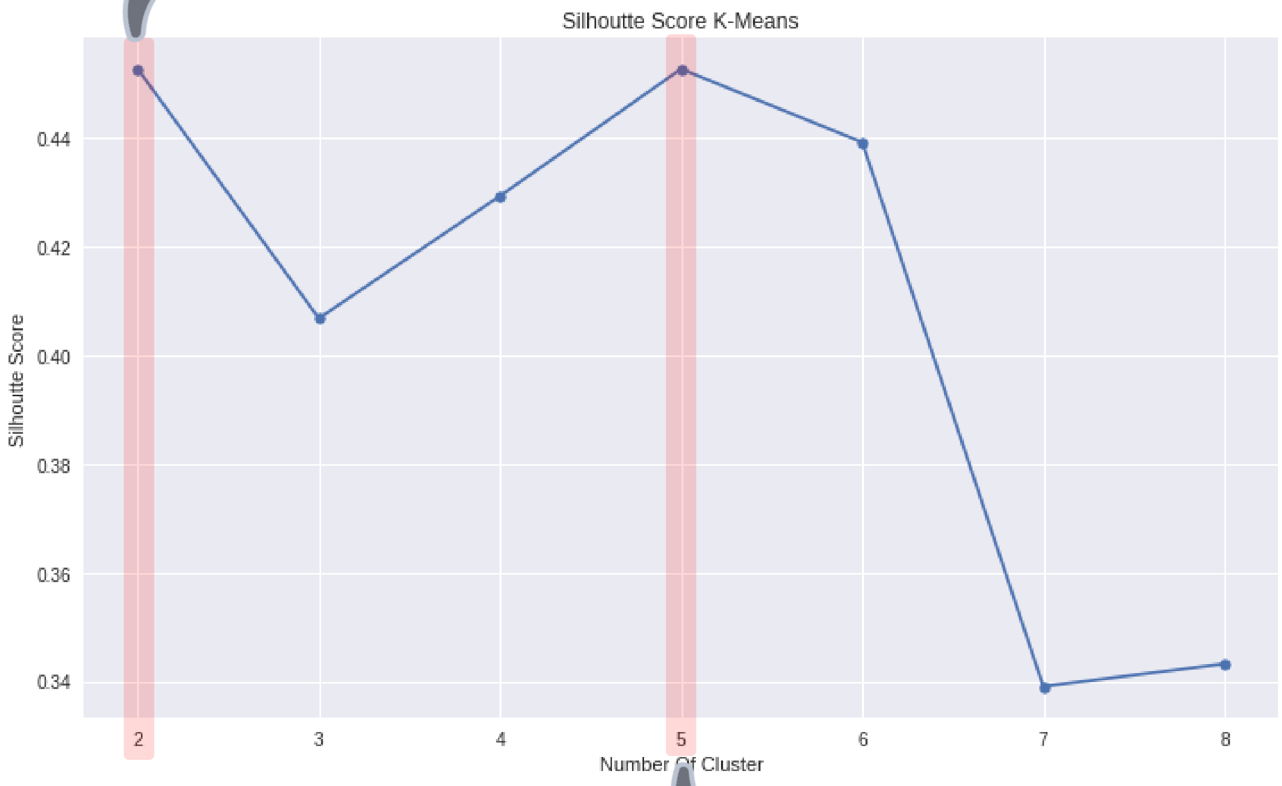
RFM Segmentation



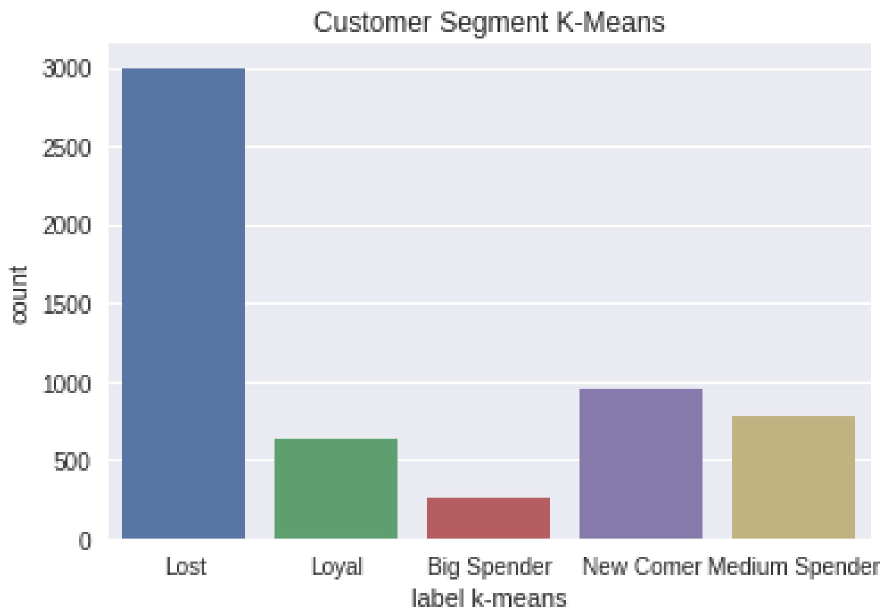


K-Means

The best
Silhouette Score

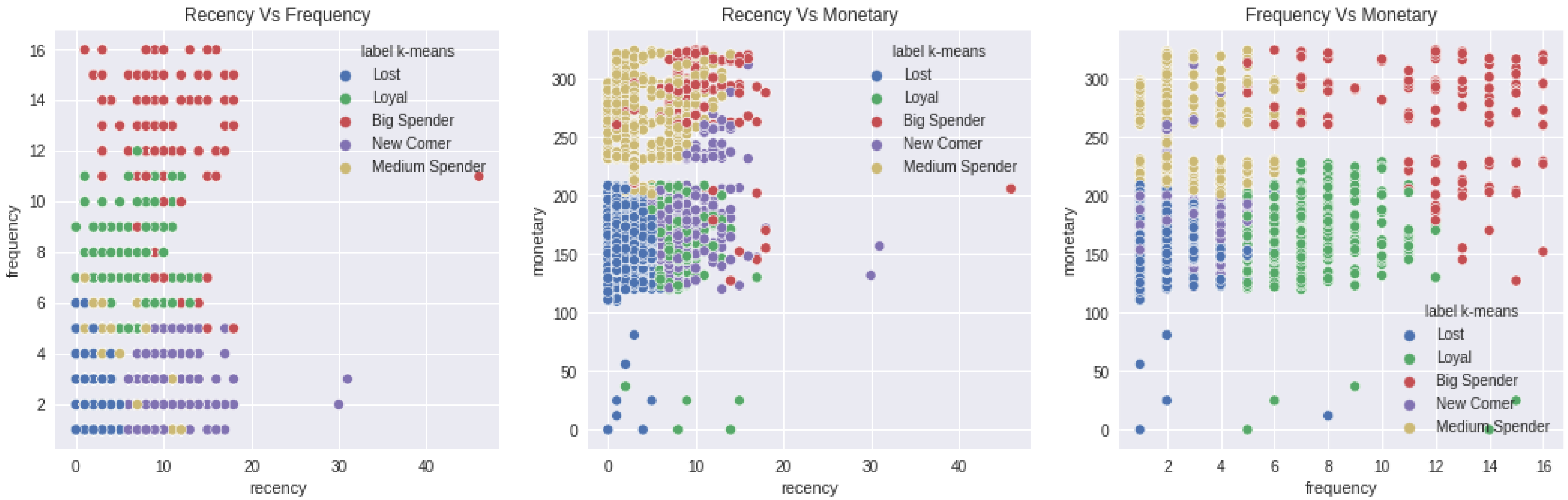


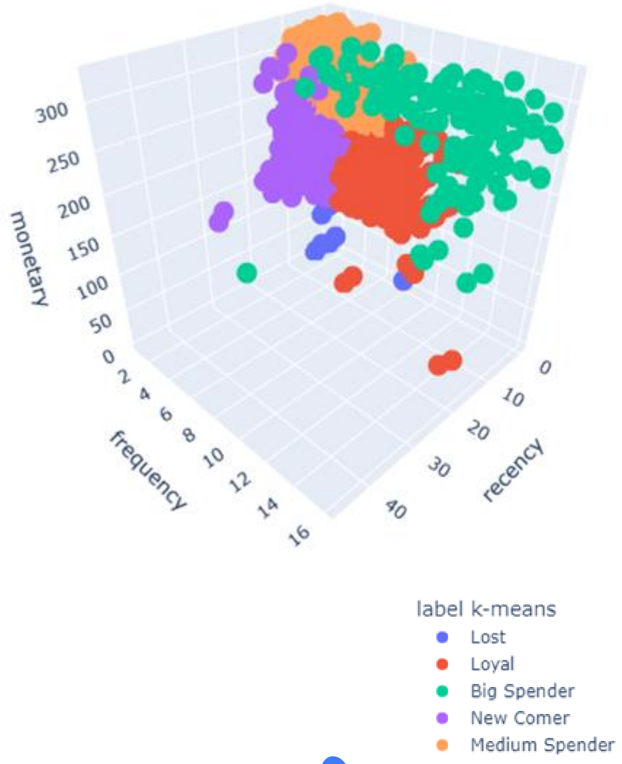
Using Clusters **5**



K-Means

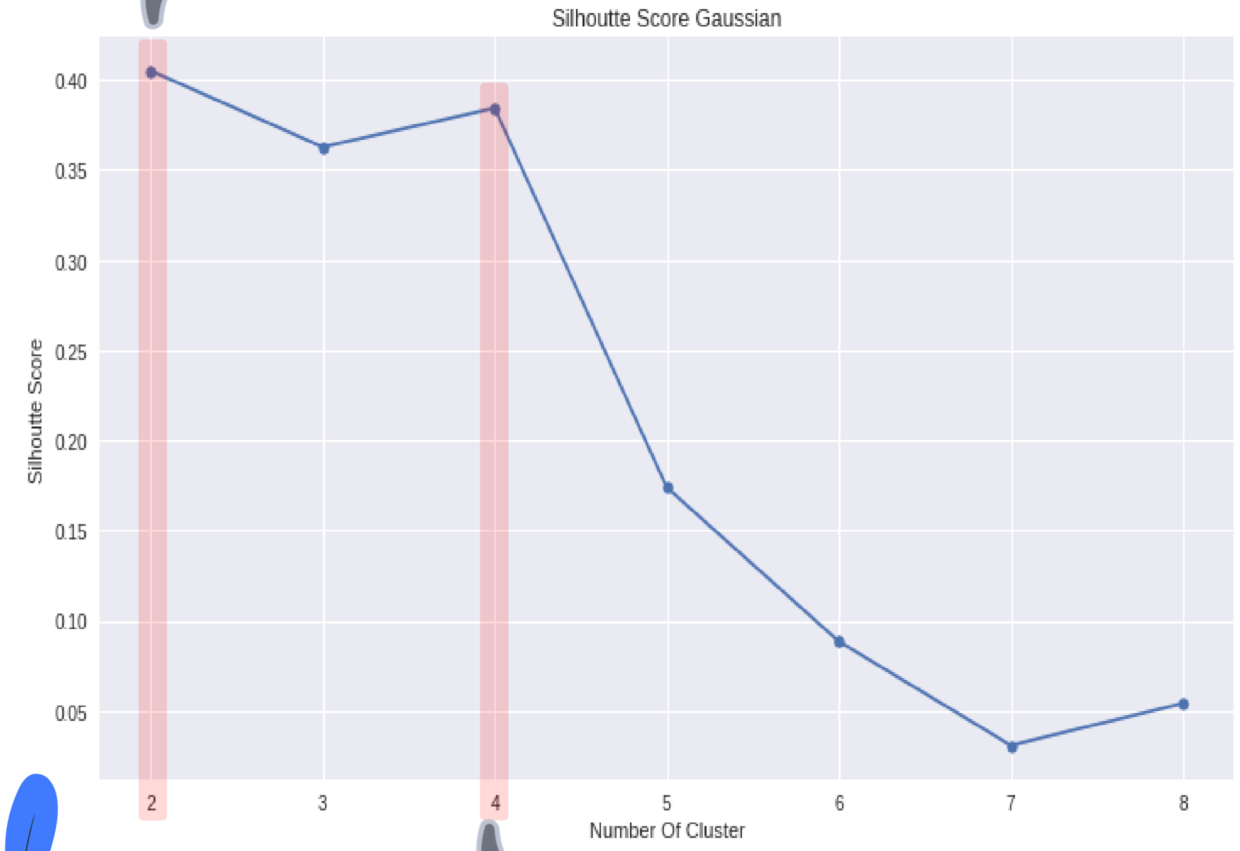
Model 2D Plot



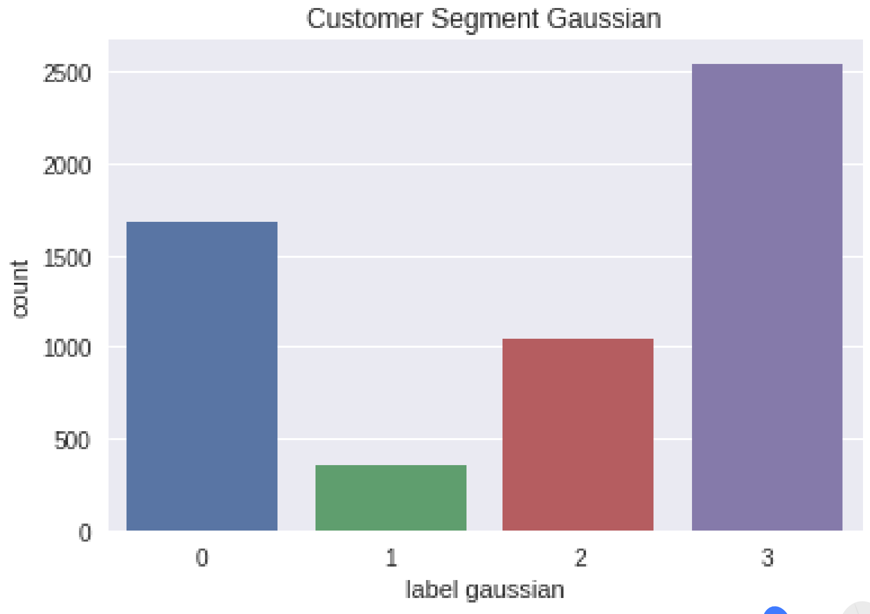


Gaussian

The best
Silhouette Score

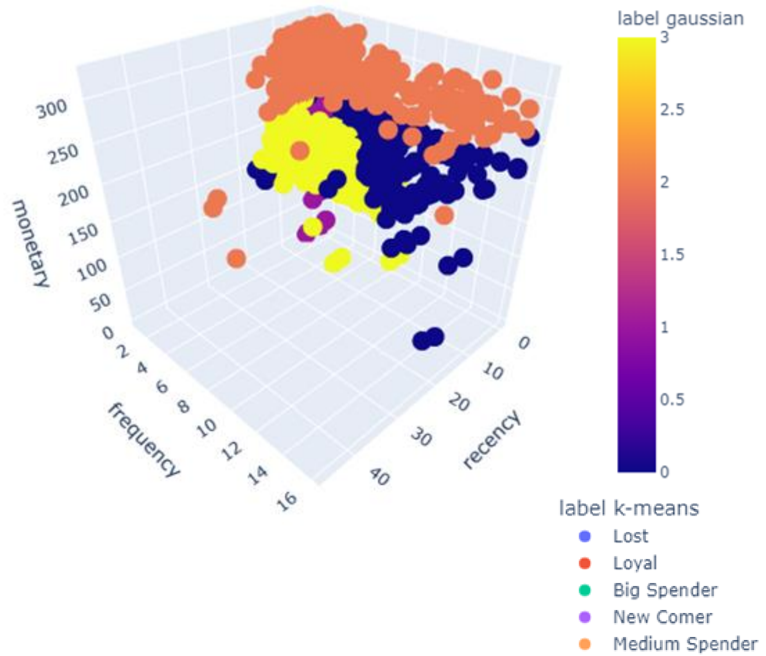


Using Clusters **4**



Gaussian

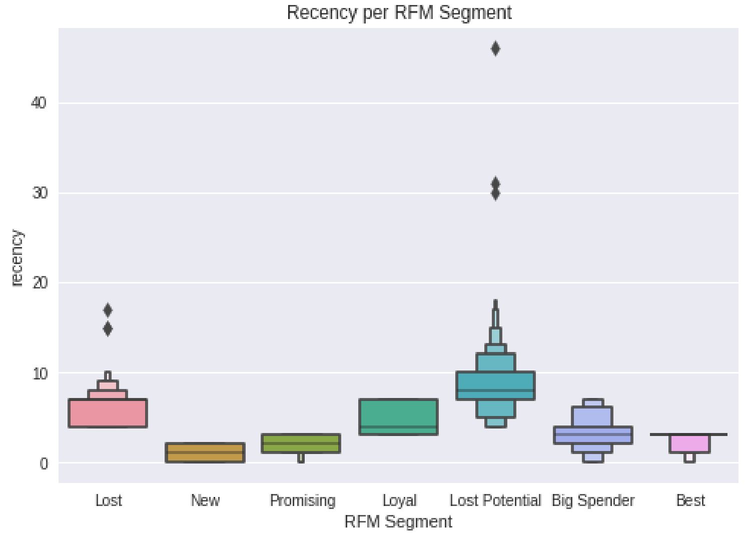
Model 2D Plot
Recency Vs Frequency
label gaussian
0
1
2
3
frequency
recency
Recency Vs Monetary
monetary
recency
Frequency Vs Monetary
monetary
frequency

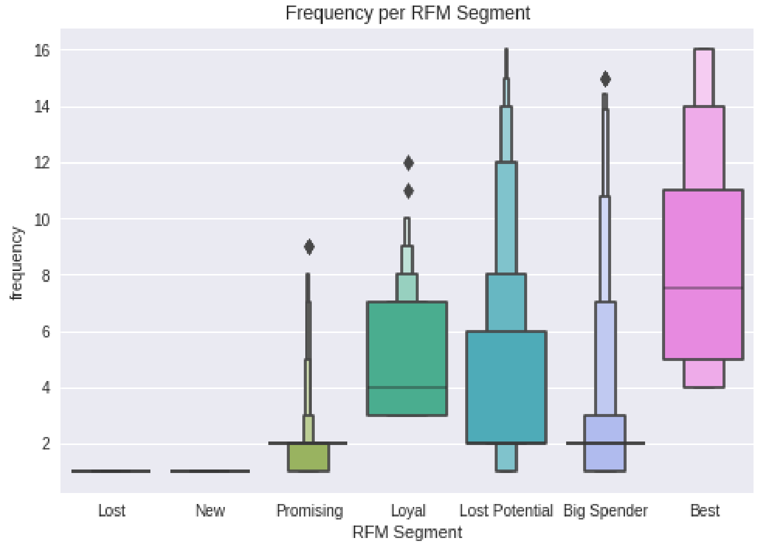


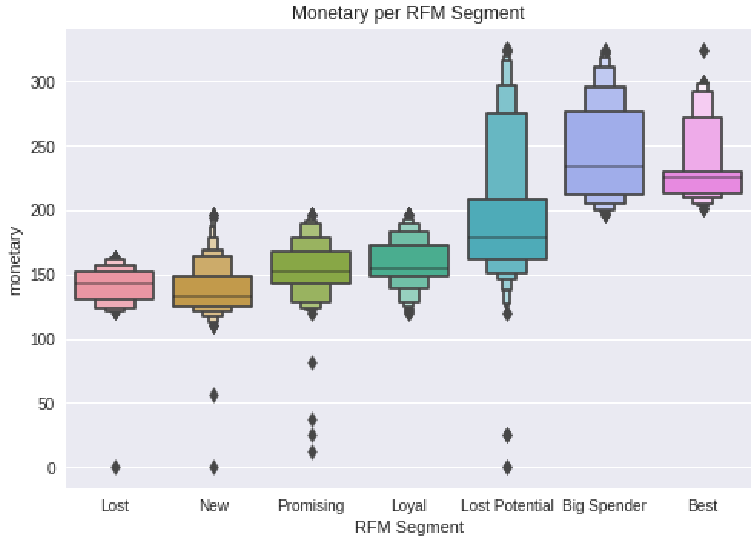
label gaussian
monetary
frequency
recency
label k-means
Lost
Loyal
Big Spender
New Comer
Medium Spender

Summary RFM Segmentation

“ The **RFM Segmentation model** is a model that has the **highest interpretation** compared to other models & this model is made with ***domain knowledge*** that we have ”





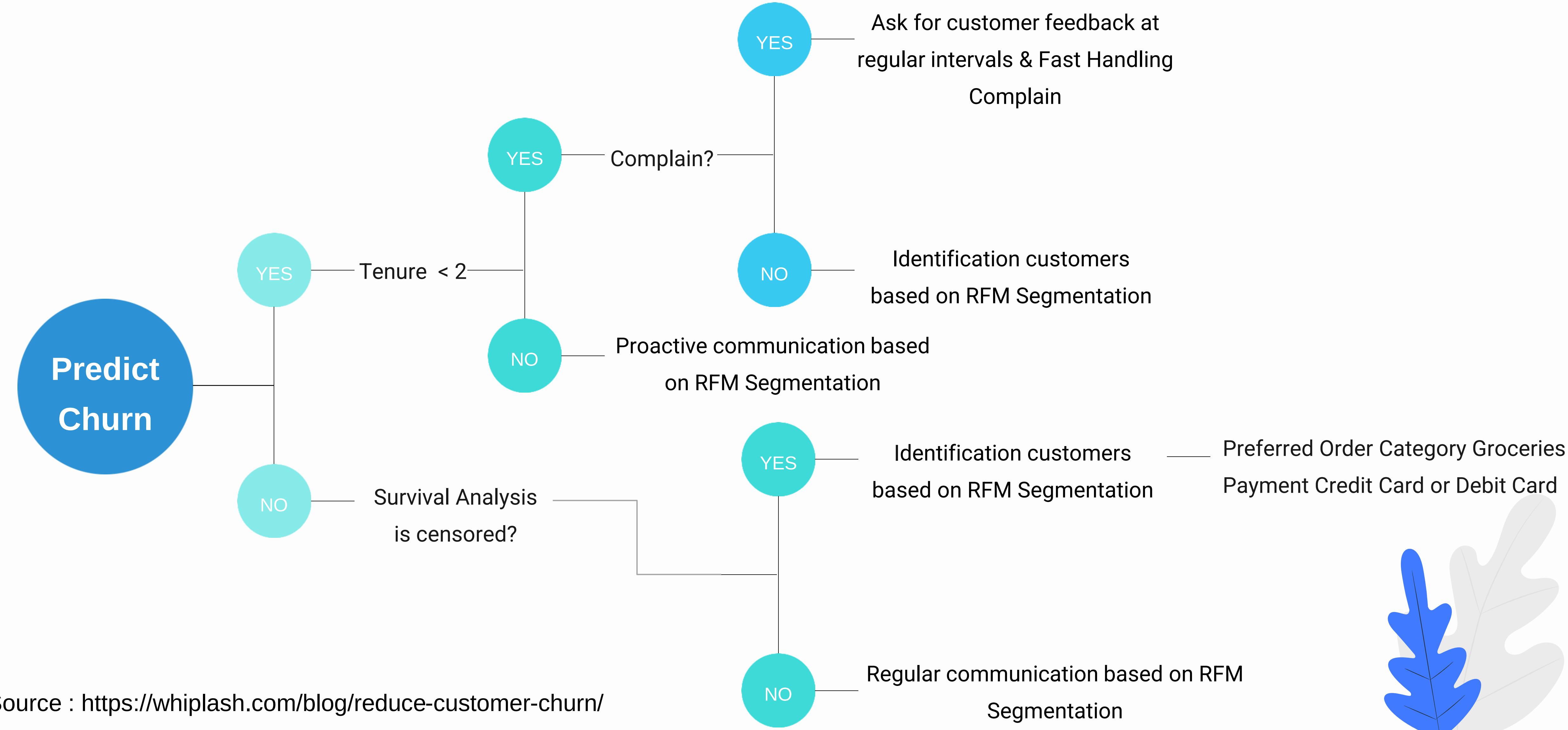


RFM Segment	Mean Recency	Mean freq	Mean monetary	Most payment type	most product buy
Best	2.625000	8.357955	230.968920	Debit Card	Fashion
Loyal	4.846307	4.842315	158.280918	Debit Card	Laptop & Acc
Big Spender	3.200843	2.567416	244.787219	Debit Card	Fashion
New	1.010135	1.000000	138.116137	Debit Card	Mobile Phone
Promising	2.079341	2.006737	153.928451	Debit Card	Mobile Phone
Lost Potential	8.461999	4.210054	195.301556	Debit Card	Laptop & Acc
Lost	6.132948	1.000000	141.281647	Credit Card	Laptop & Acc

RFM SEGMENT		STRATEGY
Best	Customers who made **transactions recently** made **frequent transactions** and had **the highest total transactions**.	Loyalty program/reward points, new product recommendations, and exclusive item offers (Cross / Up Selling Strategy).
Loyal	Customers who make **the most frequent transactions**.	Loyalty program/reward point and exclusive item offers (Cross / Up Selling Strategy).
Big Spender	Customers who have **the highest total transactions**.	Exclusive product recommendations, partnership/membership (B2B) offers, and wholesale price purchase offers (Cross / Up Selling Strategy).
New	Customers who **recently** made a **transaction** and **only made one transaction**.	Welcome email for relationship building, loyalty program/reward point offers, and discount vouchers (Cross / Up Selling Strategy).
Promising	Customers who **recently made transactions**, **as well as the frequency** and **total transactions above the average** of other customers.	Regular limited offers, discount vouchers and cashback via e-mail (Retention Strategy).
Lost Potential	Customers who have **not made transactions for a long time**, but the **frequency** and **total transactions** are **above the average** of other customers.	Regular limited offers, discount vouchers and cashback via e-mail (Retention & Reactivate Strategies)
Lost	Customers who **haven't made transactions for a long time only** made **one transaction**, and **the total** number of transactions was **small**.	Campaign via e-mail and ask for feedback. (Reactivation Strategy)

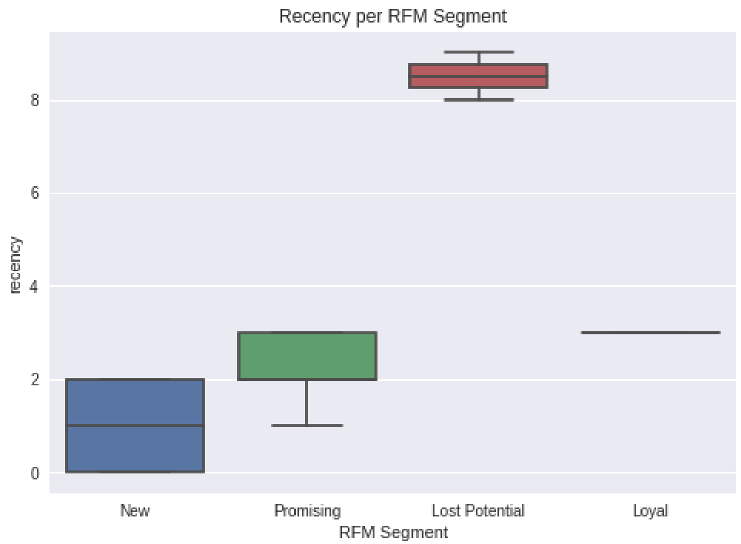
Summary RFM Segmentation

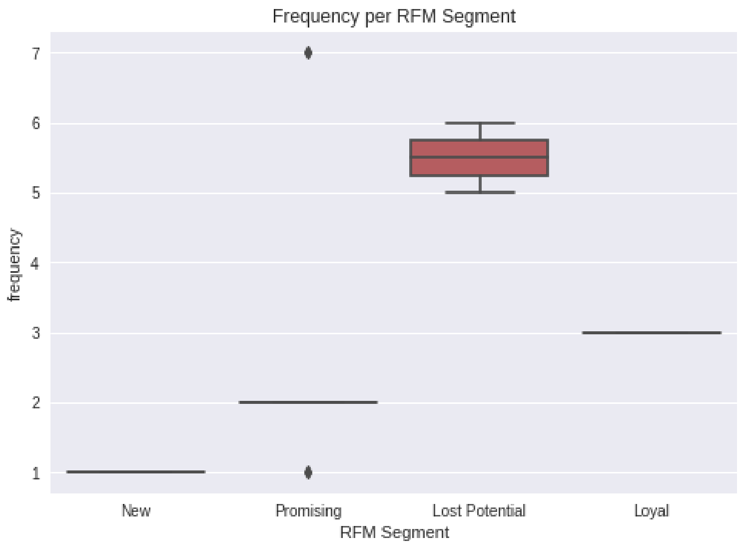
Customer Churn Treatment

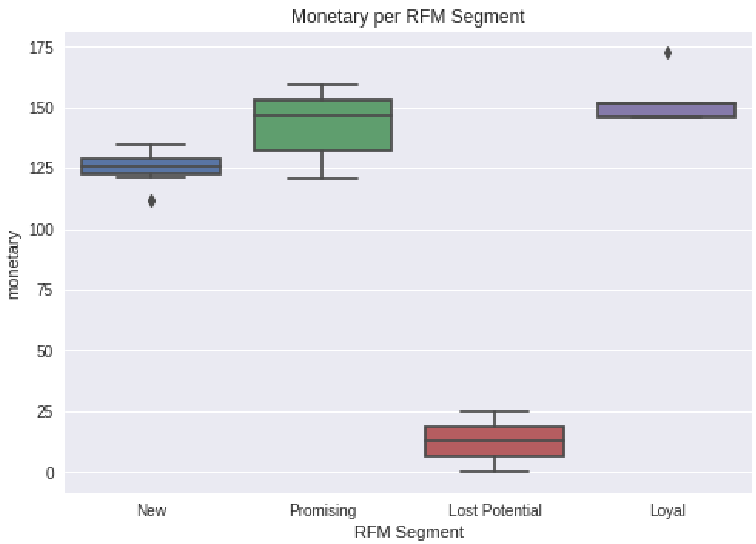


Source : https://whiplash.com/blog/reduce-customer-churn/

Censored Customer Treatment







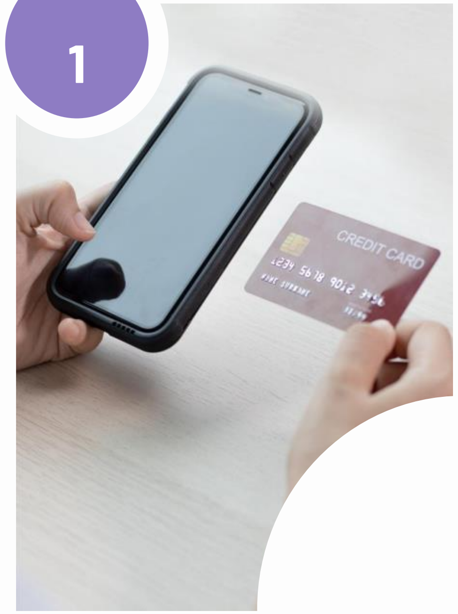
RFM Segment	RFM Segment Score	n cust	mean recency	mean freq	mean monetary	most payment type	avg review score	most product buy
Loyal	**6**	5	3.00000	3.000	153.3800	Cash on Delivery	3.80000	Mobile Phone
New	**4**	19	1.10526	1.000	125.2715	Credit Card	3.89473	Mobile Phone
Promising	**3**	23	2.30434	2.347	141.9330	Cash on Delivery	3.69565	Mobile Phone
Lost Potential	**2**	2	8.50000	5.500	12.50000	E wallet	2.00000	Mobile Phone

Recommendation Business

1

Loyal Customer

Loyalty program/reward point and **exclusive item** offers (**Cross / Up Selling** Strategy)



2

New Customer

“Welcome email” to build **relationships**, offer loyalty programs/reward points, and **discount** vouchers (**Cross / Up Selling** Strategy).

3

Promising Customer

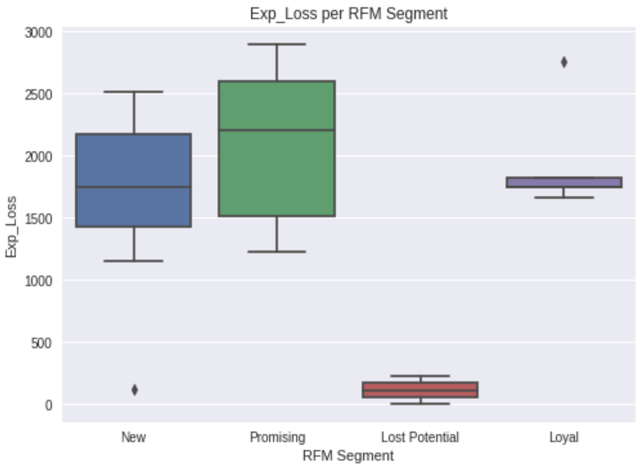
Regular **limited offers**, **discount** vouchers and **cashback** via email (Retention Strategy).

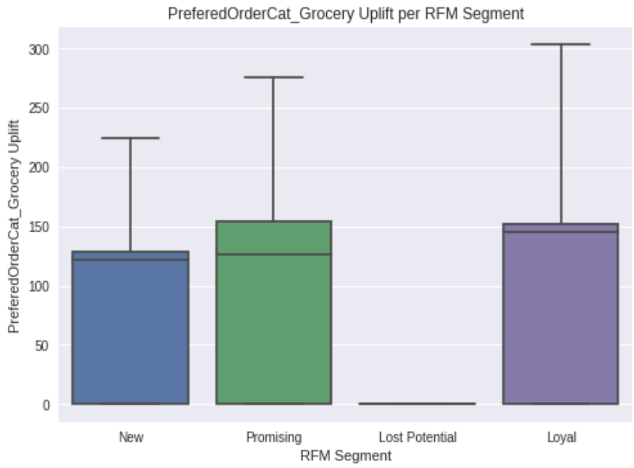
4

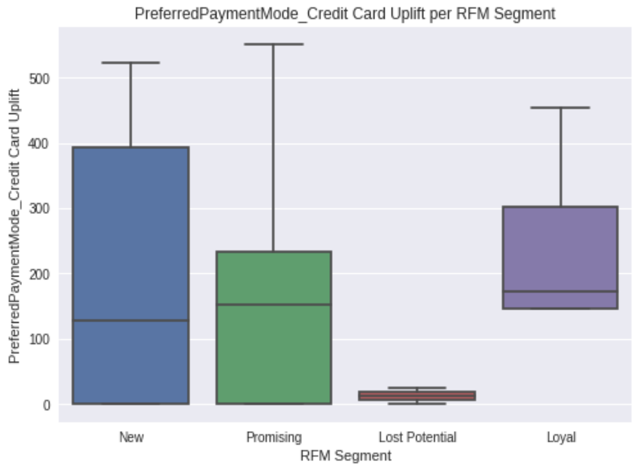
Lost Potential

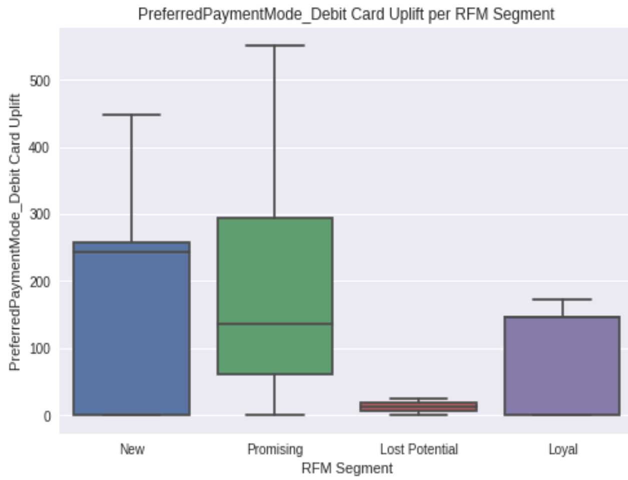
Regular **limited offers**, **discount** vouchers, and **cashback** via e-mail (Retention & Reactivate Strategies).

Estimated Loss & Revenue Uplift









$910, 687 **Total Expected Loss**

RFM Segment	RFM Segment Score	n cust	sum Exp Loss	sum Grocer Uplift	sum Credit Card Uplift	sum Debit Card Uplift
Loyal	**6**	5	9740.67	600.41	1221.61	463.76
New	**4**	19	32903.0	1372.46	3987.19	3389.94
Promising	**3**	23	48200.0	2271.93	3944.68	3975.56
Lost Potential	**2**	2	225.00	0.00	25.00	25.00

Estimated Revenue Uplift :

Order **category grocery** **$42,448**

Payment **Credit Card** **$91,785**

Payment **Debit Card** **$78,543**

SUMMARY

- From the data visualization, it is obtained that the churn ratio has a correlation with tenure, complaints, cashback Amount, & preferred order cat.
- The results of predicting churn are strongly influenced by the level of Tenure, Complaint, Number of Addresses, and cashback Amount.
- The results of the Survival Analysis, the customer has the greatest survival chance in No Complain, Marital Status Married, Payment Mode Credit Card, Order Category Grocery.

SUMMARY

RFM Segmentation results show priority customer treatment in the Loyal, New, Promising, and Lost Potential segments.

Total Expected Loss of $ 910,687

Estimated Revenue Uplift

- Order category Grocery $42,448
- Payment Credit Card $ 91,785
- Payment Debit Card $ 78,543

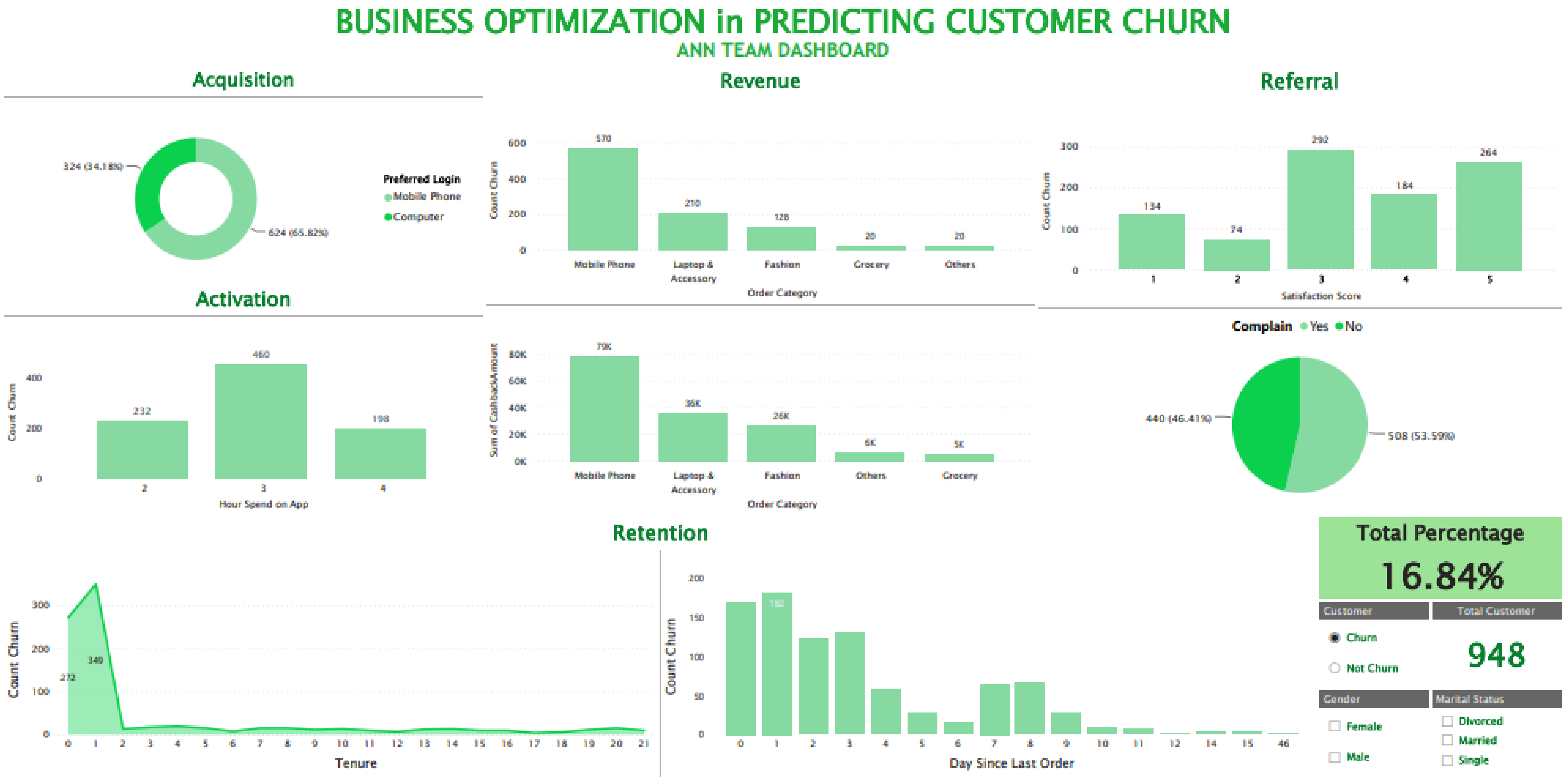
SUMMARY

- ✓ From the data visualization, it is obtained that the churn ratio has a correlation with tenure, complaints, cashback Amount, & preferred order cat.
- ✓ The results of predicting churn are strongly influenced by the level of Tenure, Complaint, Number of Addresses, and cashback Amount.
- ✓ The results of the Survival Analysis, the customer has the greatest survival chance in No Complain, Marital Status Married, Payment Mode Credit Card, Order Category Grocery.

SUMMARY

- ✓ RFM Segmentation results show priority customer treatment in the Loyal, New, Promising, and Lost Potential segments.
- ✓ Total Expected Loss of $ 910,687
- ✓ Estimated Revenue Uplift
  - Order category Grocery $42,448
  - Payment Credit Card $ 91,785
  - Payment Debit Card $ 78,543

Business Analytics Dashboard



Link : https://app.powerbi.com/groups/06be1cdf-3422-46e2-b0c1-6605fed8dfe5/reports/dcf2e131-7f8b-4661-b764-d7891b4840cb/ReportSectionc04ef661493e23894037

Bibliography

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Al-Sahaf, H., Bi, Y., Chen, Q., Lensen, A., Mei, Y., Sun, Y., Tran, B., Xue, B., & Zhang, M. (2019). A survey on evolutionary machine learning. *Journal of the Royal Society of New Zealand*, *49*(2), 205–228. https://doi.org/10.1080/03036758.2019.1609052

Apte, C. (2010). Invited Applications Paper: The Role of Machine Learning in Business Optimization. *Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010.*

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Masarifoglu, M., & Buyuklu, A. H. (2019). *Applying Survival Analysis to Telecom Churn Data* (pp. 261–275). American Journal of Theoretical and Applied Statistics.

Pinem, R. J., Afrizal, T., & Saputra, J. (2020). The Relationship of Cashback, Discount, and Voucher toward Decision to Use Digital Payment in Indonesia. *Talent Development & Excellent*, *12*(3s), 2766–2774.

Wu, X., & Meng, S. (2016). E-commerce Customer Churn Prediction Based on. *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, 1–5.

https://www.kaggle.com/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

TERIMA KASIH

Koordinator TSDN 2022: