We Care About
Your Future

E-Commerce Customer Churn Prediction

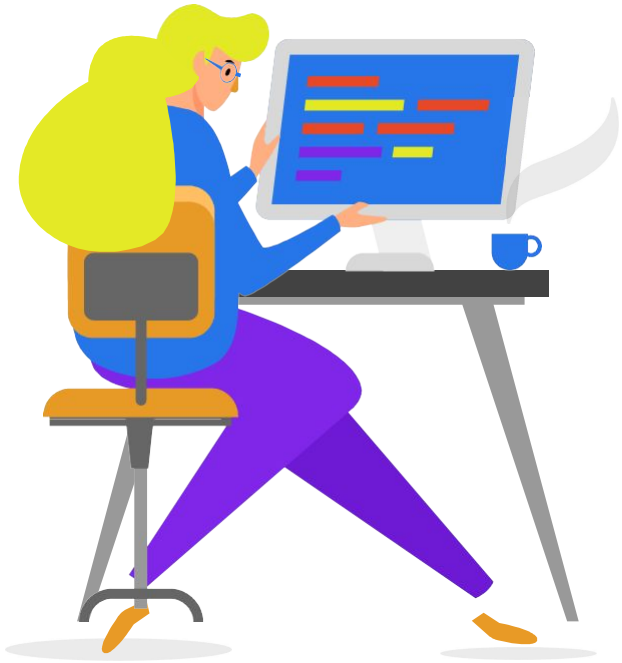
Final Project

Oleh :

1. M Aldi
2. Elka Mustika
3. Fauzan Ihza Fajar
4. Saprina Hani Haqyah

We Care About
Your Future

Table of Contents



We Care About
Your Future

01 Business Understanding

Saat ini, perkembangan e-commerce di Indonesia memang terus meningkat.

Kondisi customer churn adalah hal yang mempengaruhi perkembangan bisnis. Salah satu mekanisme terbaik untuk mempertahankan pelanggan saat ini adalah mengidentifikasi potensi churn dan merespons dengan cepat untuk mencegahnya.

How ?

Teknik data mining dapat diterapkan Untuk menganalisis perilaku pelanggan Dan untuk memprediksi pengurangan Pelanggan potensial sehingga strategi Pemasaran khusus dapat diadopsi untuk mempertahankannya.

We Care About
Your Future

01 Business Understanding

Business Question

1. Bagaimana persentase customer yang memilih churn?
2. Keadaan apa saja yang mempengaruhi churn rate?

Goal

1.Menjelaskan penyebab potensi terjadinya churn rate berdasarkan keadaan customer.
2.Membangun machine learning untuk mendeteksi customer yang akan churn.

02 Data Preprocessing

Variable	Discerption
CustomerID	Unique customer ID
Churn	Churn Flag
Tenure	Tenure of customer in organization
PreferredLoginDevice	Preferred login device of customer
CityTier	City tier
WarehouseToHome	Distance in between warehouse to home of customer
PreferredPaymentMode	Preferred payment method of customer
Gender	Gender of customer
HourSpendOnApp	Number of hours spend on mobile application or website
NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
PreferedOrderCat	Preferred order category of customer in last month
SatisfactionScore	Satisfactory score of customer on service
MaritalStatus	Marital status of customer
NumberOfAddress	Total number of added added on particular customer
Complain	Any complaint has been raised in last month
OrderAmountHikeFromlastYear	Percentage increases in order from last year
CouponUsed	Total number of coupon has been used in last month
OrderCount	Total number of orders has been places in last month
DaySinceLastOrder	Day Since last order by customer
CashbackAmount	Average cashback in last month

5630 rows ×20 columns

We Care About
Your Future

02 Data Preprocessing

Handling Missing Values

```
CustomerID                     0
Churn                          0
Tenure                       264
PreferredLoginDevice           0
CityTier                       0
WarehouseToHome              251
PreferredPaymentMode           0
Gender                         0
HourSpendOnApp               255
NumberOfDeviceRegistered       0
PreferedOrderCat               0
SatisfactionScore              0
MaritalStatus                  0
NumberOfAddress                0
Complain                       0
OrderAmountHikeFromlastYear  265
CouponUsed                   256
OrderCount                   258
DaySinceLastOrder            307
CashbackAmount                 0
dtype: int64
```

before

df = df.fillna(df.median())


```
Churn                        0
Tenure                       0
PreferredLoginDevice         0
CityTier                     0
WarehouseToHome              0
PreferredPaymentMode         0
Gender                       0
HourSpendOnApp               0
NumberOfDeviceRegistered     0
PreferedOrderCat             0
SatisfactionScore            0
MaritalStatus                0
NumberOfAddress              0
Complain                     0
OrderAmountHikeFromlastYear  0
CouponUsed                   0
OrderCount                   0
DaySinceLastOrder            0
CashbackAmount               0
dtype: int64
```

After

Karena terdapat nilai ekstrem pada data(outliers), maka untuk menangani missing values tersebut Kita akan menggunakan nilai median. Median merupakan nilai terbaik untuk nilai imputasi karena robust terhadap outliers.

02 Data Preprocessing

Handling Inconsistent Data

```
for col in df.columns:
        print("=="*10)
        print(f" {col}", df[col].unique())
```

```
PreferredLoginDevice ['Mobile Phone' 'Phone' 'Computer']

PreferredPaymentMode ['Debit Card' 'UPI' 'CC' 'Cash on Delivery' 'E wallet' 'COD' 'Credit Card']

PreferedOrderCat ['Laptop & Accessory' 'Mobile' 'Mobile Phone' 'Others' 'Fashion' 'Grocery']
```

Karena terdapat inconsistent data pada beberapa feature
Maka perlu ada perubahan sehingga merujuk pada kategori yang sama.

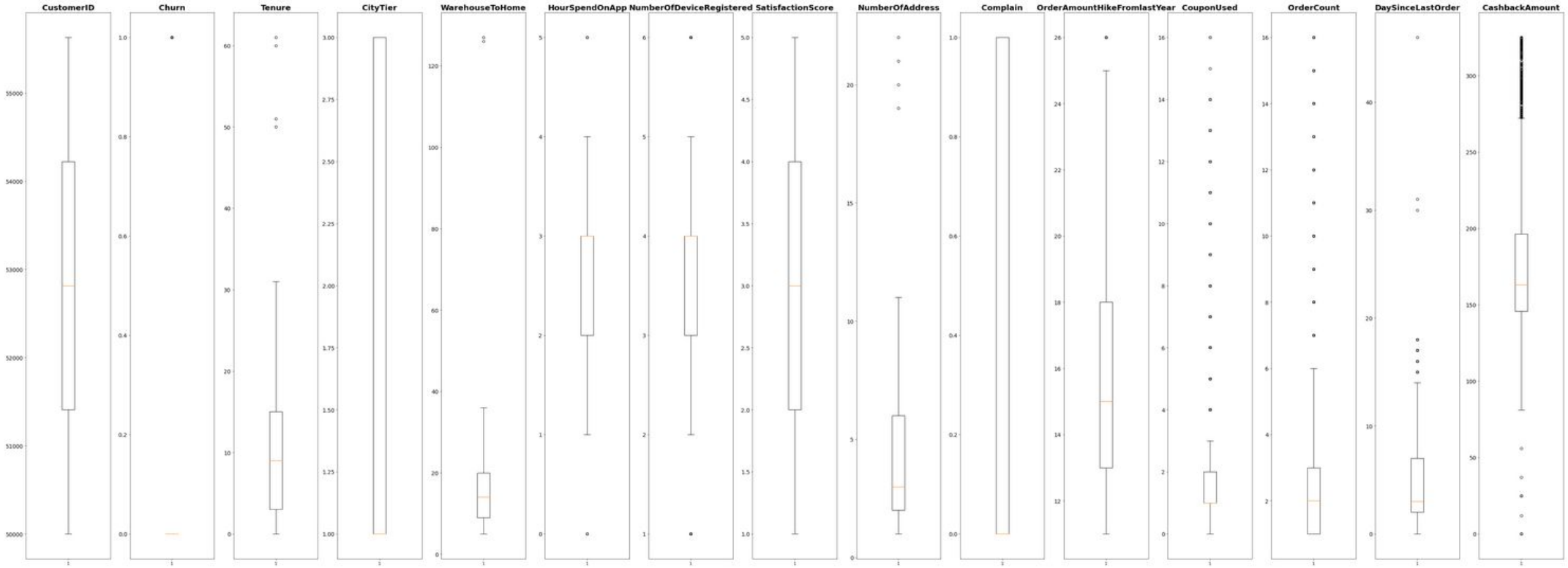
Mobile = Mobile phone
Phone = Mobile Phone
CC = Credit Card
COD = Cash on Delivery

```
df= df.replace(['Mobile','Phone','CC','COD'],['Mobile Phone','Mobile Phone','Credit Card','Cash on Delivery'])
```

We Care About
Your Future

02 Data Preprocessing

Handling Handling Outliers : MetodeZ-Score

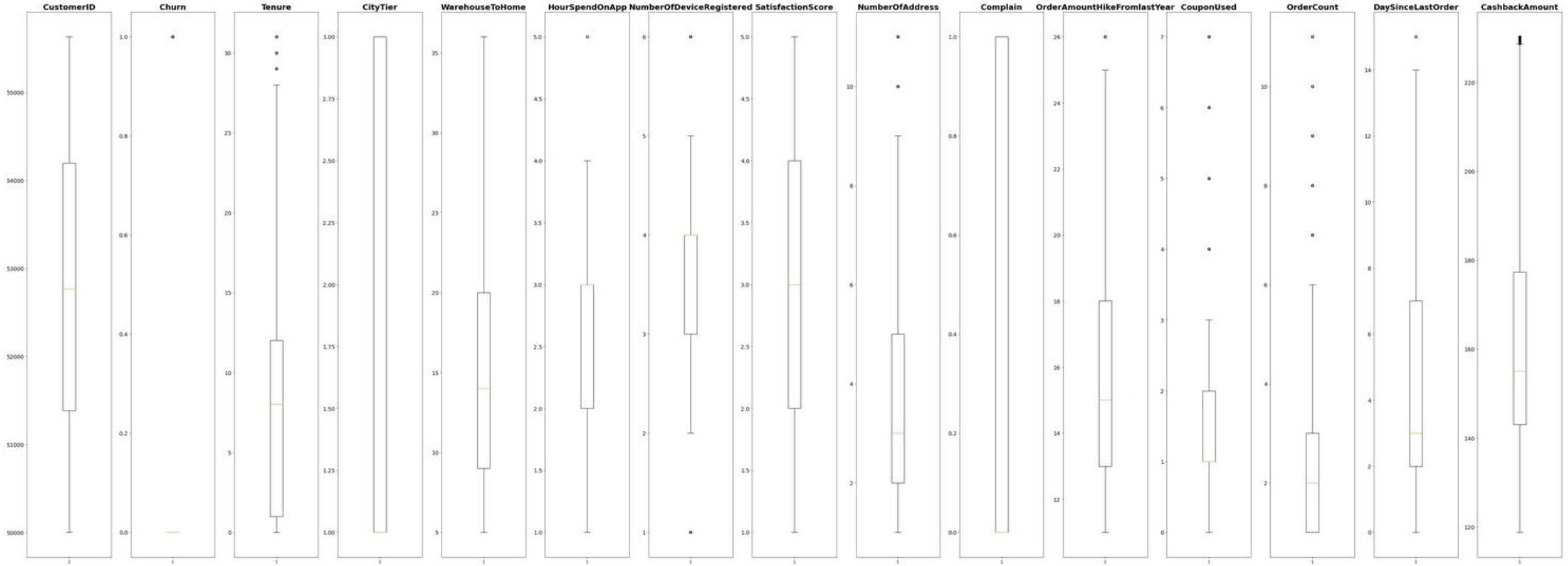


Pada kasus ini, untuk menangani data outliers menggunakan metode Z-Score.
Setelah data outliers dihilangkan,dimensi data berkurang. Metode Z-Score cocok digunakan karena data outliers yang dihilangkan tidak terlalu banyak.

We Care About
Your Future

02 Data Preprocessing

Handling Handling Outliers : MetodeZ-Score

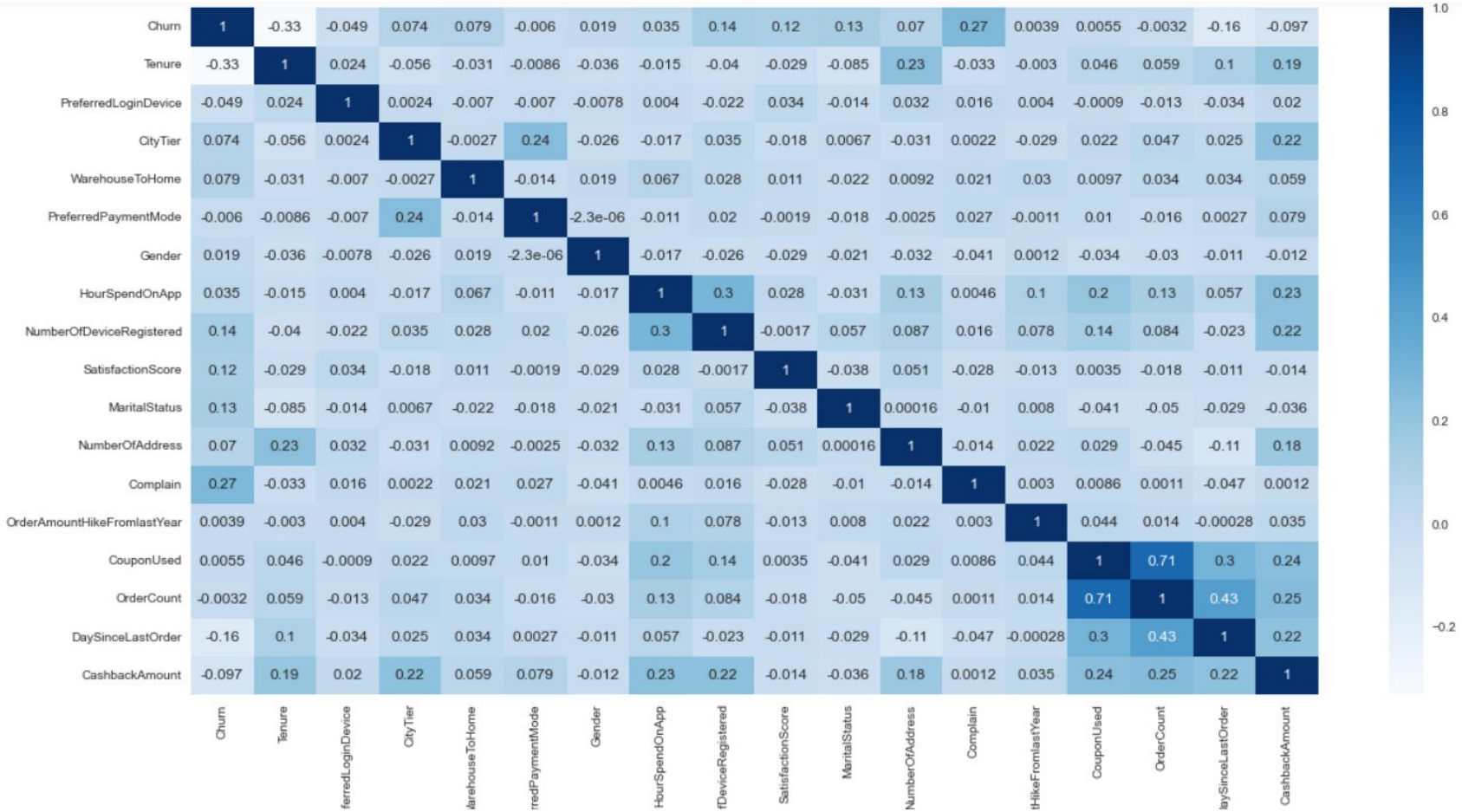


Data shape sebelum di hilangkan outliers (5630, 20)
Data shape setelah handling outliers (4588, 20)
Threshold pada metode zscore sebesar 3 dan 1.2 untuk kolom “cashbackamount”

We Care About
Your Future

02 Data Preprocessing

Heat Map



Menggunakan Data Heat Map untuk menentukan feature mana yang memiliki korelasi yang tinggi dan analisa dari tim kami

02 Data Preprocessing

Feature Engineering

•Remove Feature : CustomerID, Tenure,Hour Spend App, Number of device registered, Number of address, order amount hike from last year, Coupon used, order Count, Day since last order , cashback amount
•Feature Encoding : Label Encoding
•Feature Scaling : Robust Scaling

Resampling Dataset

Separate train and test set with 82.60% train set and 18.40% test set

Target imbalanced, Oversampling using SMOTE to make target to be balanced

We Care About
Your Future

03 Exploratory Data Analysis

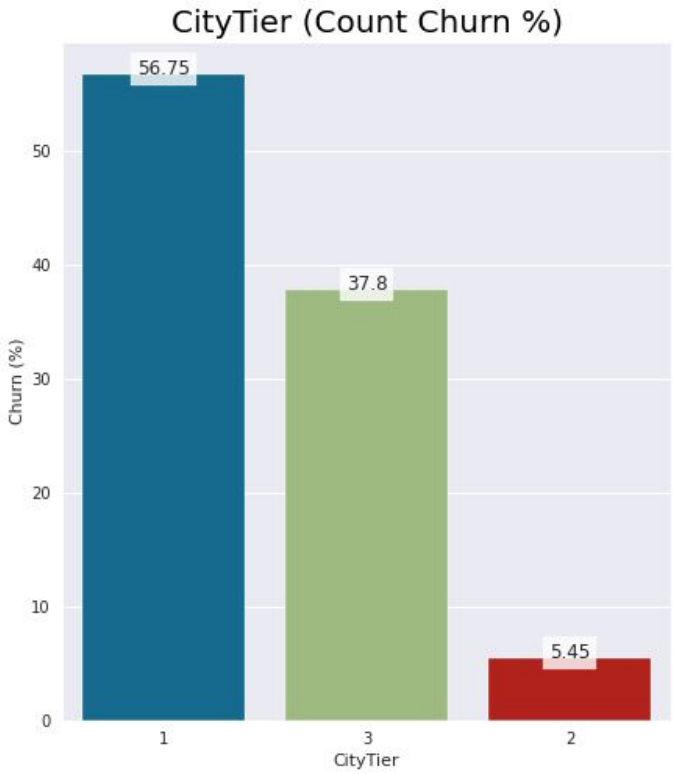
Sekitar 18.40% dari total 4588 customer yang tercatat memilih untuk churn sedangkan 81.60% memilih tidak churn.

We Care About
Your Future

03 Exploratory Data Analysis

INSIGHT

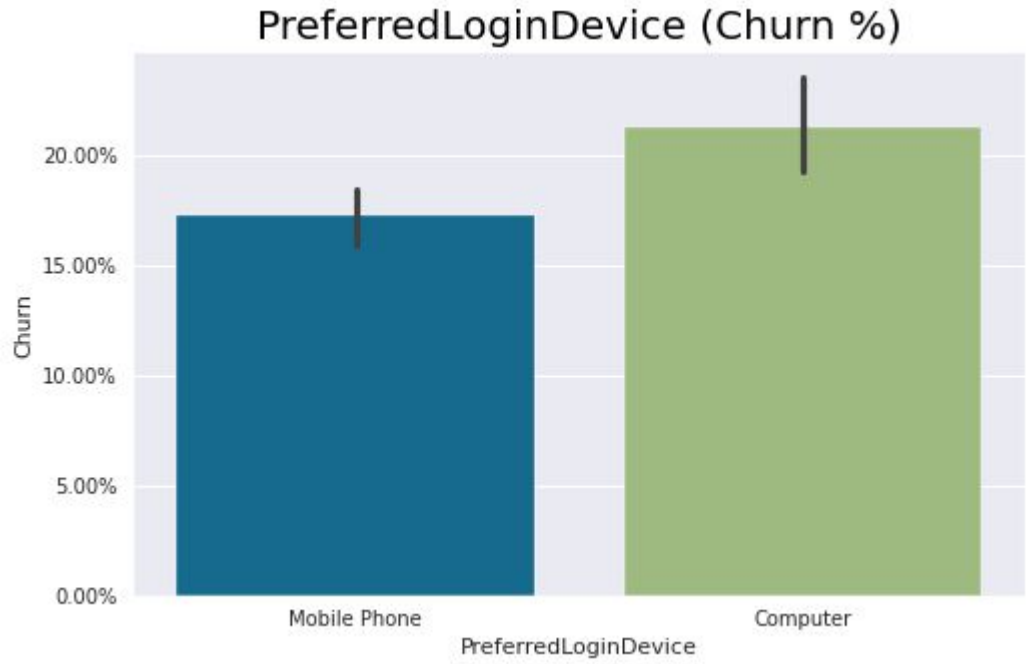
City tier atau tingkatan kota
semakin tinggi tingkatan kota semakin tinggi pula potensi unutk churn. Namun khusus untuk kasus ini tingkatan kota 3 lebih rendah dikarenakan populasi di city tier 3 jauh lebih banyak ketimbang city tier 2.



INSIGHT

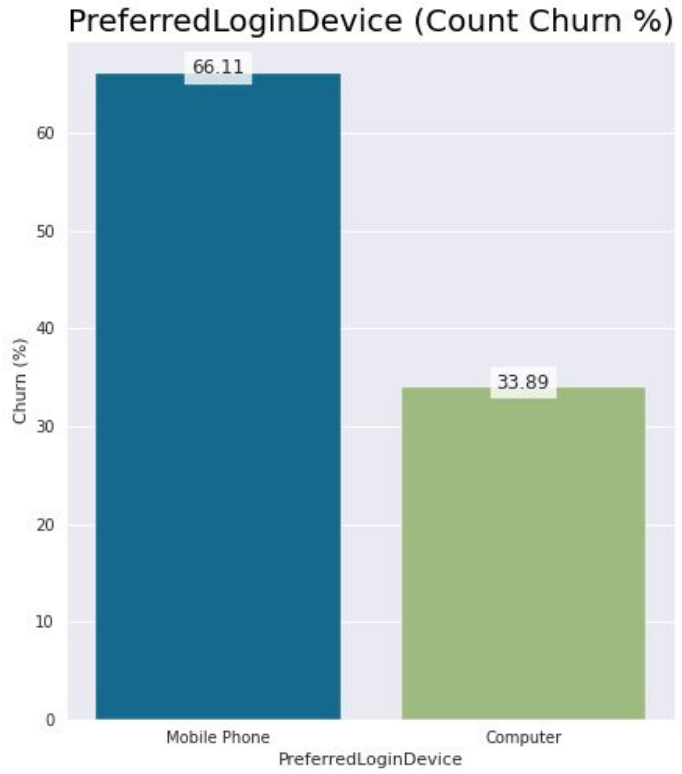
City tier atau tingkatan kota
berdasarkan data churn City Tier 1 adalah tingkatan kota tertinggi/ metropolitan menunjukkan tingkat churn tertinggi yaitu 56.75%

03 Exploratory Data Analysis



INSIGHT

preferred login device menunjukkan bahwa penggunaan login device computer memiliki tingkat churn yang tinggi

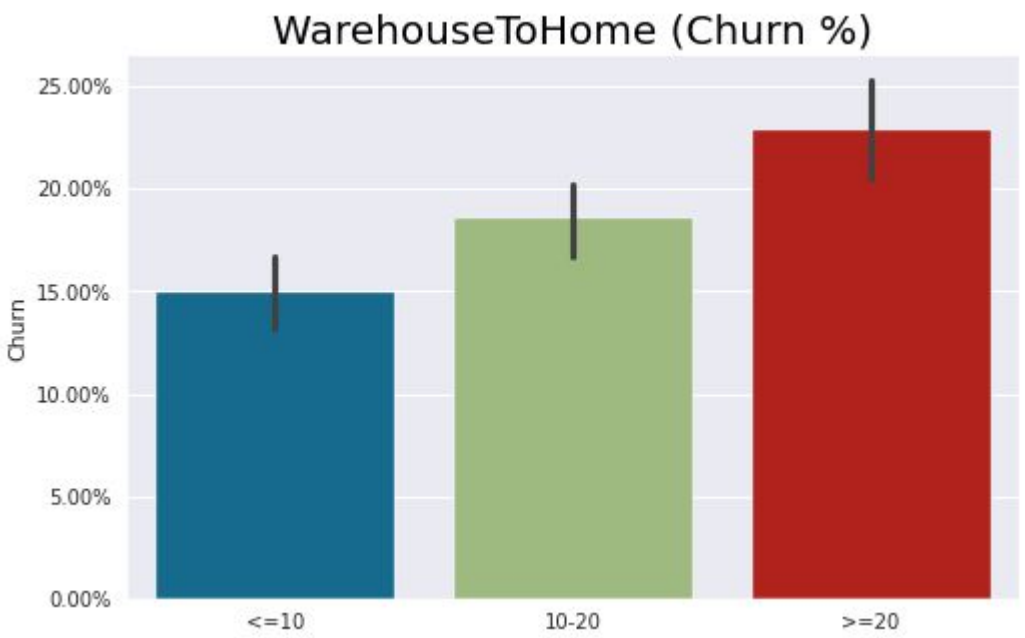


INSIGHT

khusus berdasarkan data churn saja preferred login device menunjukkan bahwa penggunaan login device phone memiliki tingkat churn yang tinggi yaitu 66.11%.

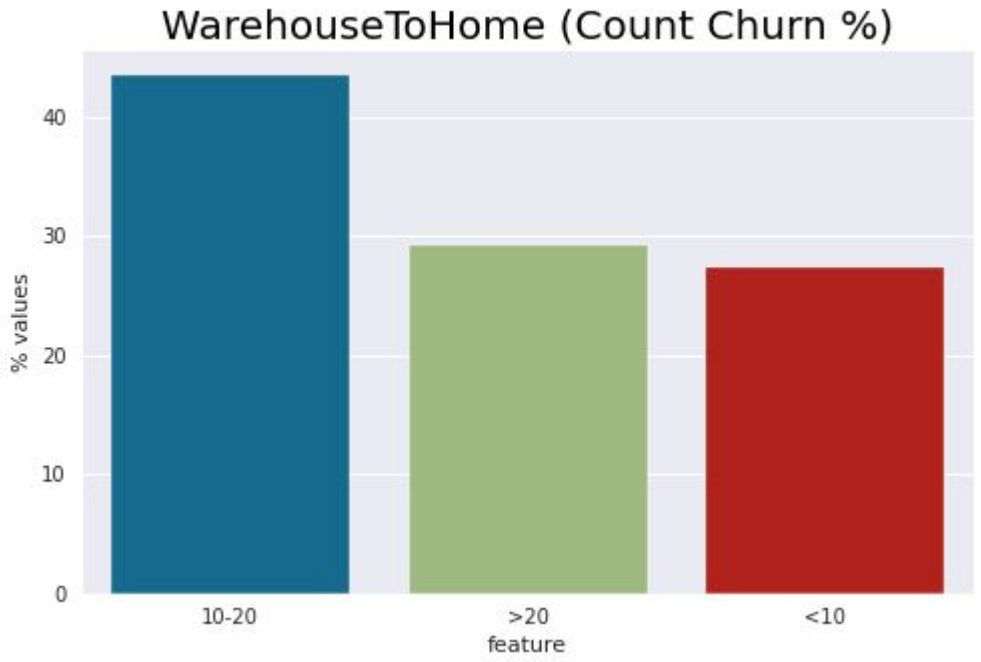
We Care About
Your Future

03 Exploratory Data Analysis



INSIGHT

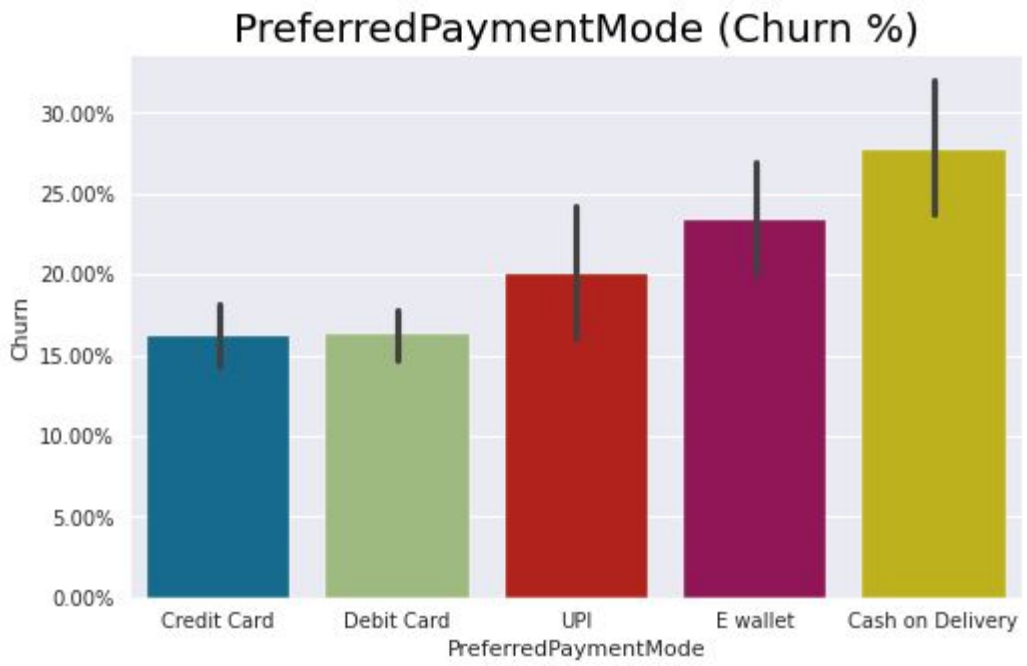
Dari data di atas terlihat bahwa semakin jauh jarak rumah customer semakin tinggi persentasi untuk churn. Bisa dikarenakan oleh ongkos kirim makin mahal dengan semakin jauh lokasi tempat tinggal dari warehouse



INSIGHT

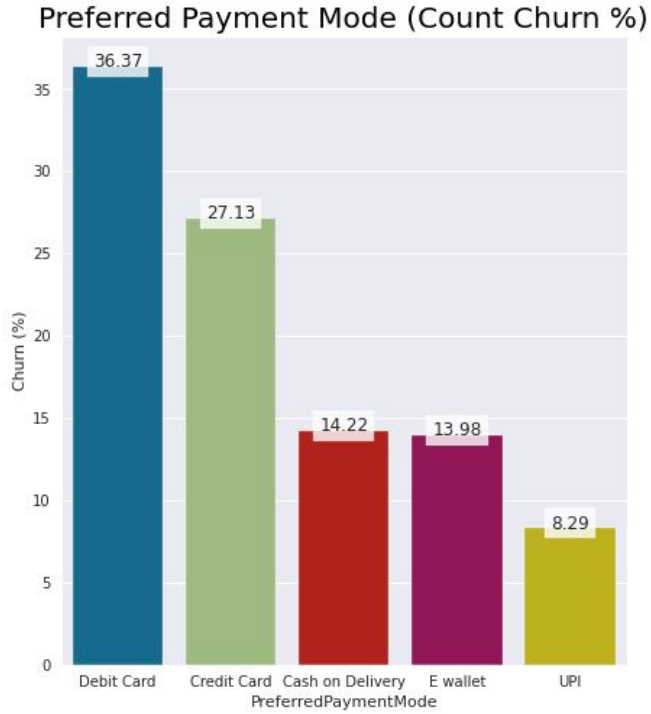
Dari data khusus churn ini menunjukkan lebih dari 40% lebih banyak churn adalah yang jarak 10-20.

03 Exploratory Data Analysis



INSIGHT

Preferred Payment Mode menunjukkan bahwa motode cash on delivery yang berpotensi churn karena bisa disebabkan karena ketidakpuasan pelanggan saat melihat paket yang sampai tidak sesuai atau terjadi kerusakan.

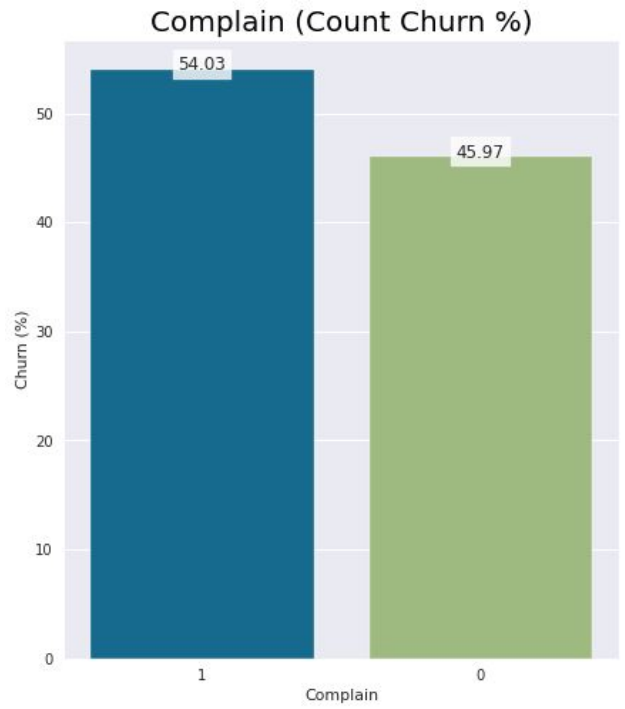


INSIGHT

Preferred Payment Mode khusus untuk data churn menunjukkan debit card paling tinggi yaitu 36,37%

We Care About
Your Future

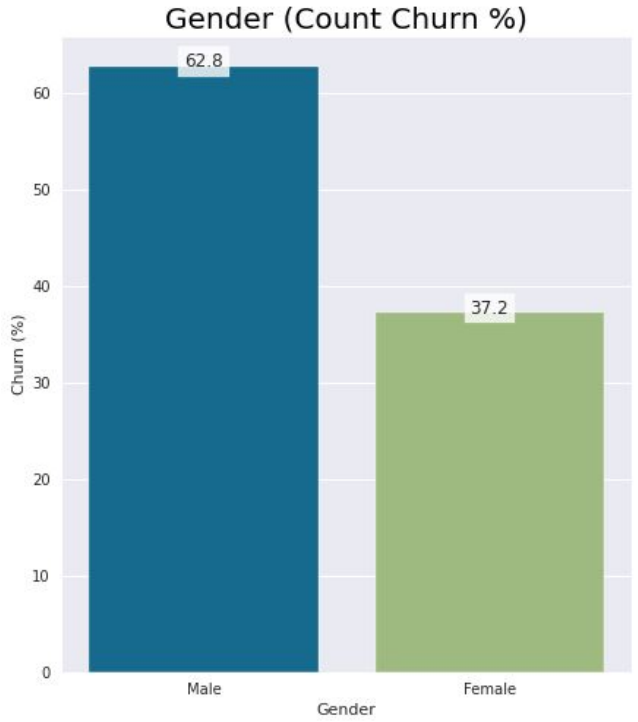
03 Exploratory Data Analysis

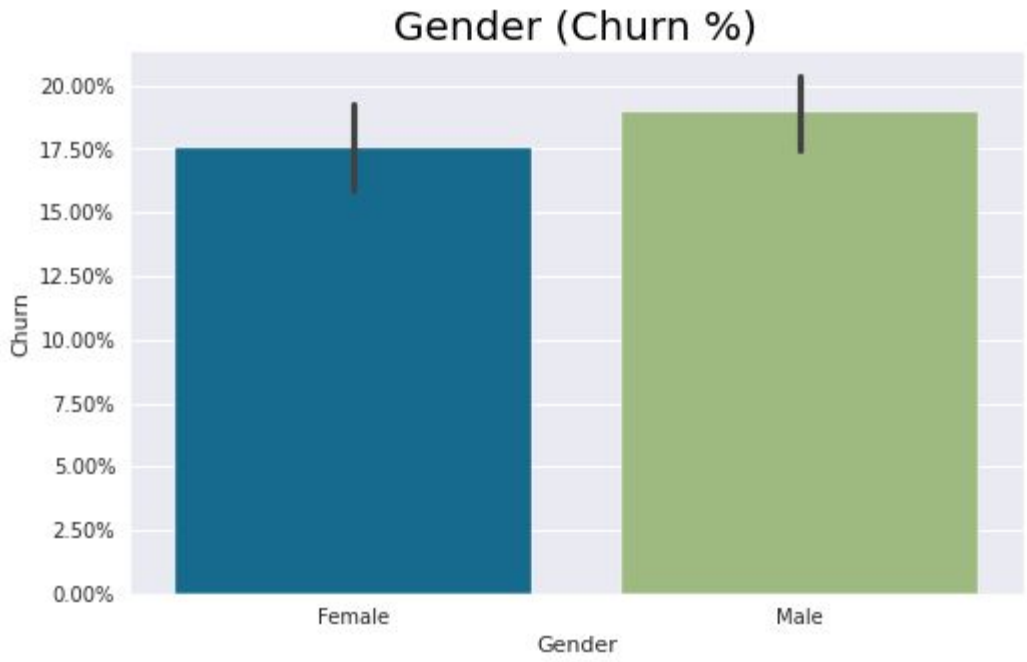


INSIGHT

Complain customer mempengaruhi terhadap churn. Untuk kasus complain customer harus segera ditanggapi dengan baik agar bisa mencegah churn di kemudian hari.

03 Exploratory Data Analysis



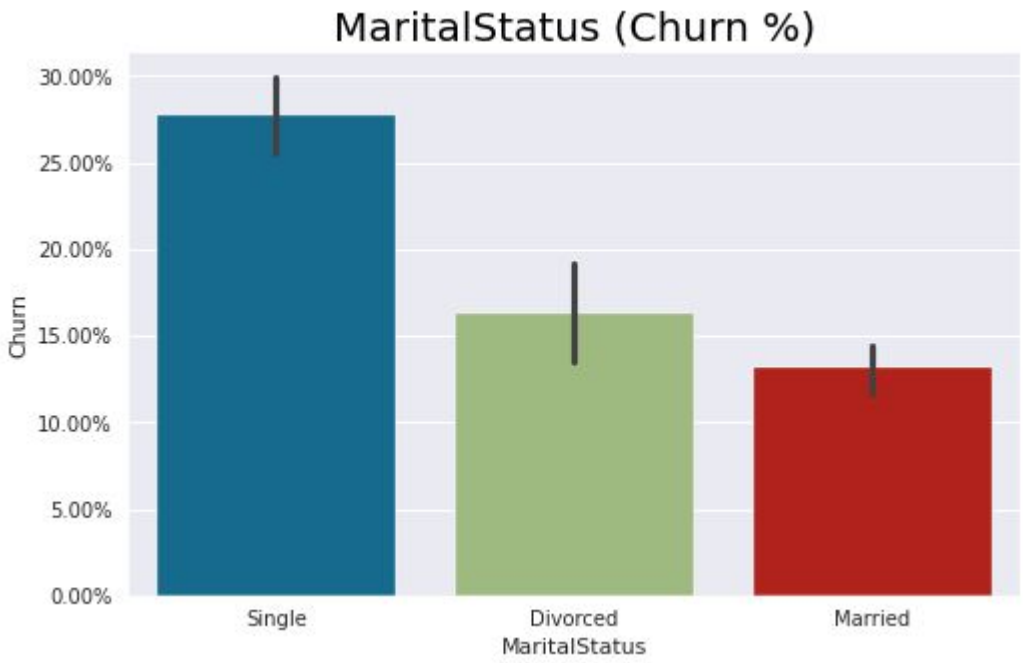


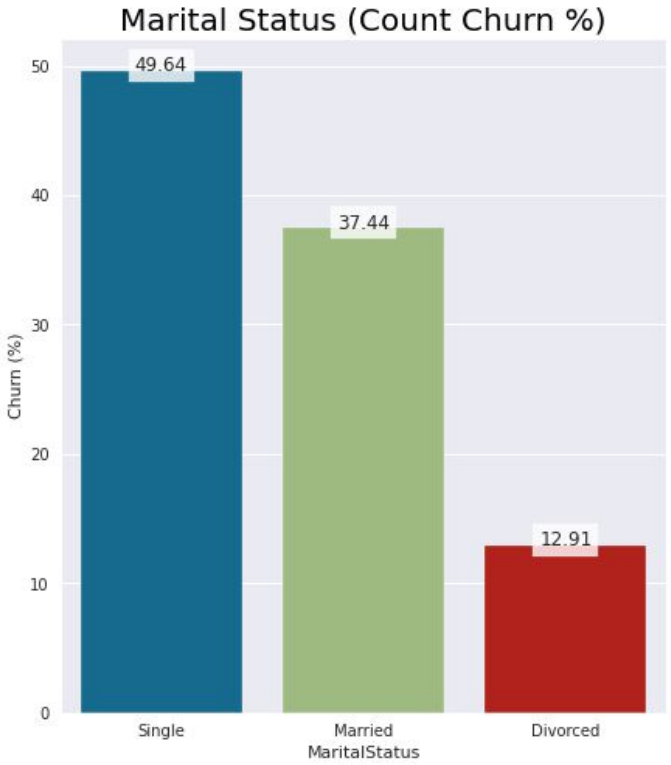
INSIGHT

Terlihat pada data bahwa customer berjenis kelamin laki-laki lebih berpotensi untuk churn sedangkan pada customer berjenis kelamin perempuan tingkat churn lebih rendah

We Care About
Your Future

03 Exploratory Data Analysis

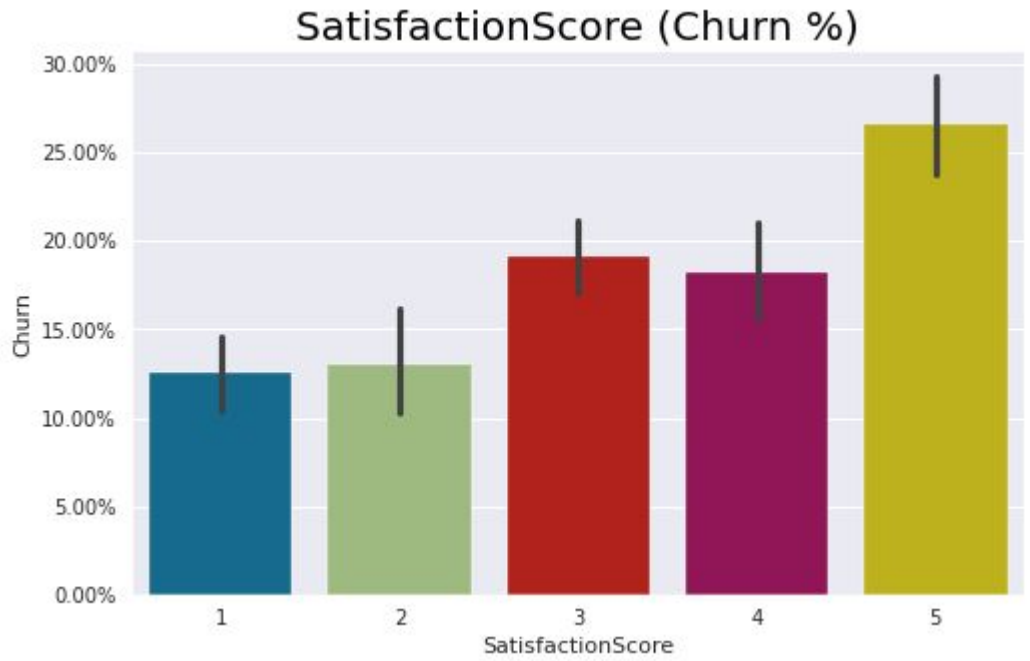


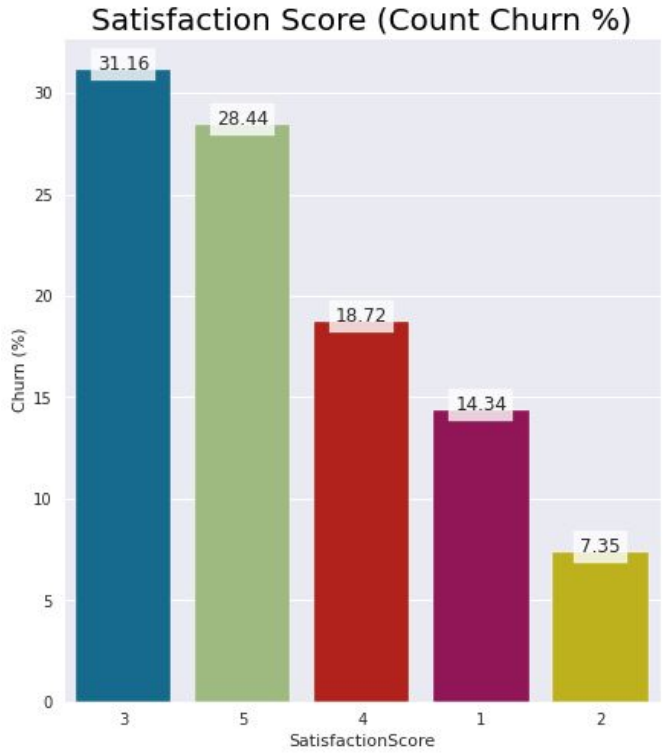


INSIGHT

Customer dengan status single memiliki tingkat churn yang tinggi dibandingkan customer dengan status married dan divorce.

03 Exploratory Data Analysis





INSIGHT

Data SatisfactionScore dipengaruhi oleh beberapa faktor salah satunya pelayanan ke customer atau kondisi produknya atau kecepatan pengiriman.

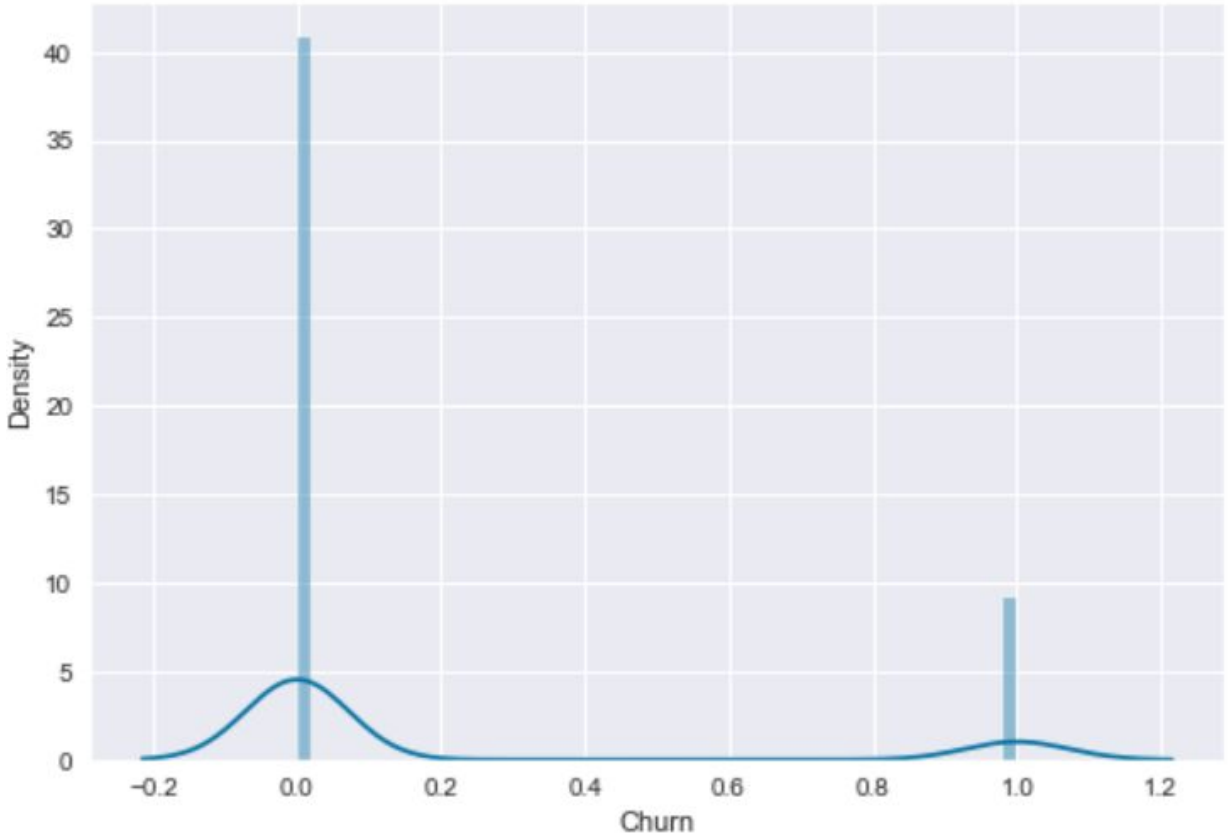
We Care About
Your Future

04 Model and Evaluation

1. Model machine learning yang dibuat dengan membandingkan beberapa model machine learning yaitu ExtraTrees, Random Forest, Decision Tree, XGBoost, KNN, Logistic Regresion, SVM, Naïve Bayer, MLP, AdaBoost, Bagging, GradientBoosting.
2. Metode Evaluasi machine learning yang digunakan yaitu cross validation
3. Metric evaluasi yang digunakan adalah accuracy, precision,recall dan f-1 score.

We Care About
Your Future

04 Modelling



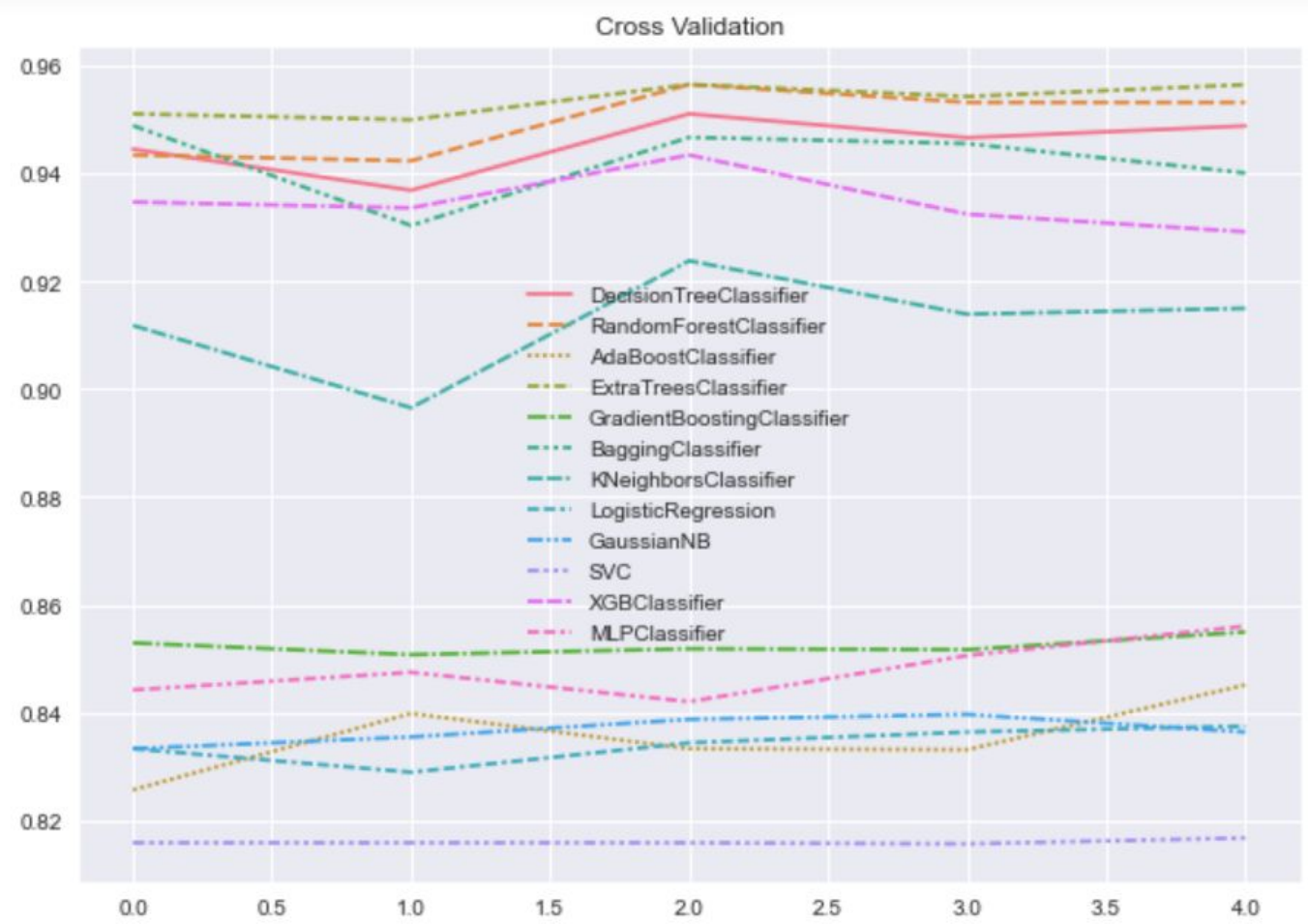
0	2795
1	646

Oversampling

1	2795
0	2795

We Care About
Your Future

04 Modelling (Cross Validation)



	score_mean	score_std
ExtraTreesClassifier	0.953575	0.002709
RandomForestClassifier	0.949653	0.005727
DecisionTreeClassifier	0.945511	0.004862
BaggingClassifier	0.942241	0.006642
XGBClassifier	0.934610	0.004748
KNeighborsClassifier	0.912163	0.008829
GradientBoostingClassifier	0.852442	0.001438
MLPClassifier	0.848084	0.004929
GaussianNB	0.836749	0.002285
AdaBoostClassifier	0.835442	0.006607
LogisticRegression	0.834134	0.002966
SVC	0.816042	0.000384

04 Modelling (Cross Validation)

	Model	fit_time	score_time	test_accuracy	test_precision	test_recall	test_f1
0	ExtraTreesClassifier	0.357341	0.040867	0.953140	0.926318	0.810425	0.864254
1	RandomForestClassifier	0.409632	0.027683	0.948127	0.905407	0.802141	0.850365
2	DecisionTreeClassifier	0.012497	0.009374	0.944422	0.895240	0.791448	0.839569
3	BaggingClassifier	0.069154	0.006248	0.943331	0.886929	0.794992	0.837937
4	XGBClassifier	0.627420	0.018831	0.934610	0.891560	0.734545	0.804672
5	KNeighborsClassifier	0.018827	0.047203	0.912163	0.878370	0.606629	0.717560
6	GradientBoostingClassifier	0.310543	0.009373	0.852442	0.792332	0.270132	0.402062
7	MLPClassifier	12.247384	0.018123	0.845901	0.785455	0.232213	0.355559
8	GaussianNB	0.009374	0.006249	0.836749	0.710906	0.190786	0.300303
9	AdaBoostClassifier	0.165532	0.024994	0.835442	0.662874	0.214490	0.324057
10	LogisticRegression	0.041030	0.003126	0.834134	0.724811	0.158798	0.260132
11	SVC	0.370300	0.046866	0.816042	0.000000	0.000000	0.000000

04 Modelling (Uji Nilai Akurasi)

	train score	test score	difference
ExtraTreesClassifier	0.984973	0.935484	0.049489
RandomForestClassifier	0.984973	0.931997	0.052977
DecisionTreeClassifier	0.984973	0.915432	0.069542
BaggingClassifier	0.982469	0.913688	0.068781
XGBClassifier	0.973703	0.907585	0.066118
KNeighborsClassifier	0.944723	0.863121	0.081602
MLPClassifier	0.874597	0.804708	0.069890
GradientBoostingClassifier	0.839356	0.787271	0.052085
SVC	0.801431	0.750654	0.050777
AdaBoostClassifier	0.741145	0.723627	0.017518
LogisticRegression	0.672987	0.696600	0.023612
GaussianNB	0.667442	0.707934	0.040492

Berdasarkan proses crosvalidasi dan uji akurasi diperoleh model terbaik yaitu ExtraTreesClassifier dengan tingkat akurasi:

1. Trainingg Accuracy : 98.5 %
2. Test Accuracy : 93.2 %

04 Model Evaluation (Confusion Matrix)

	precision	recall	f1-score	support
0	0.95	0.97	0.96	949
1	0.83	0.76	0.79	198
accuracy			0.93	1147
macro avg	0.89	0.86	0.88	1147
weighted avg	0.93	0.93	0.93	1147

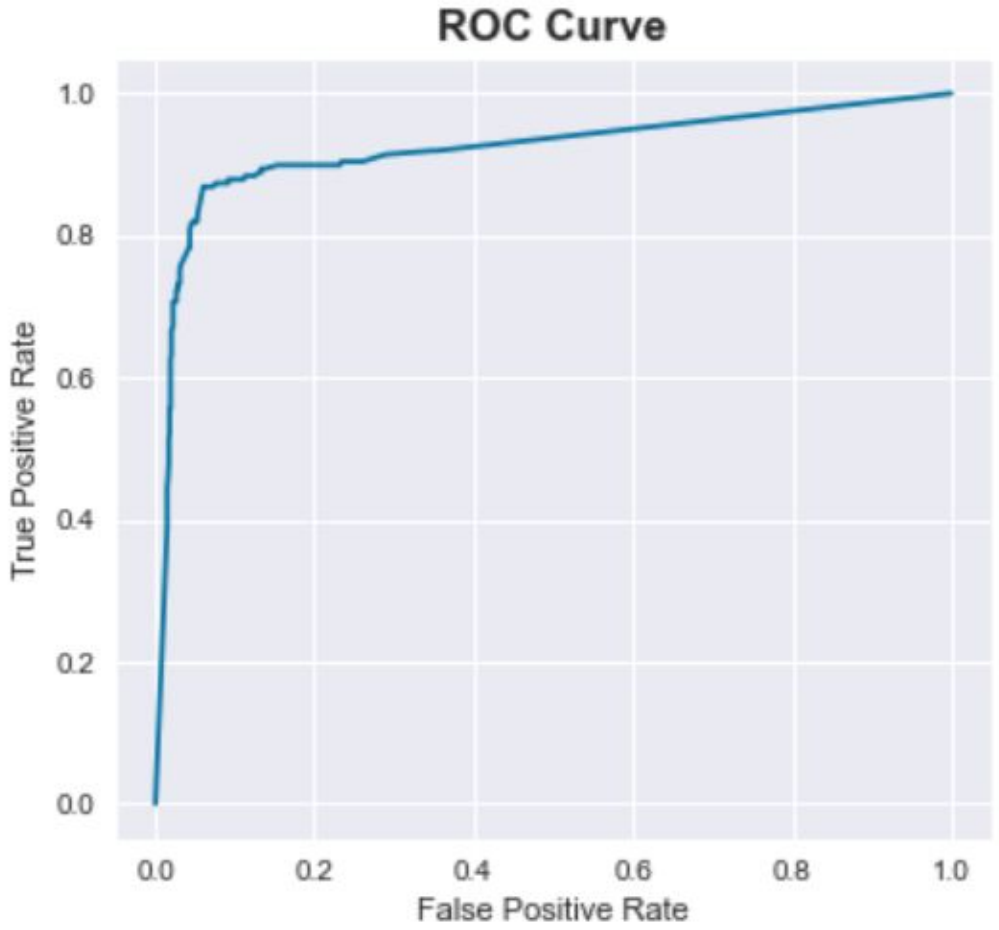
True Positives(TP) = 919

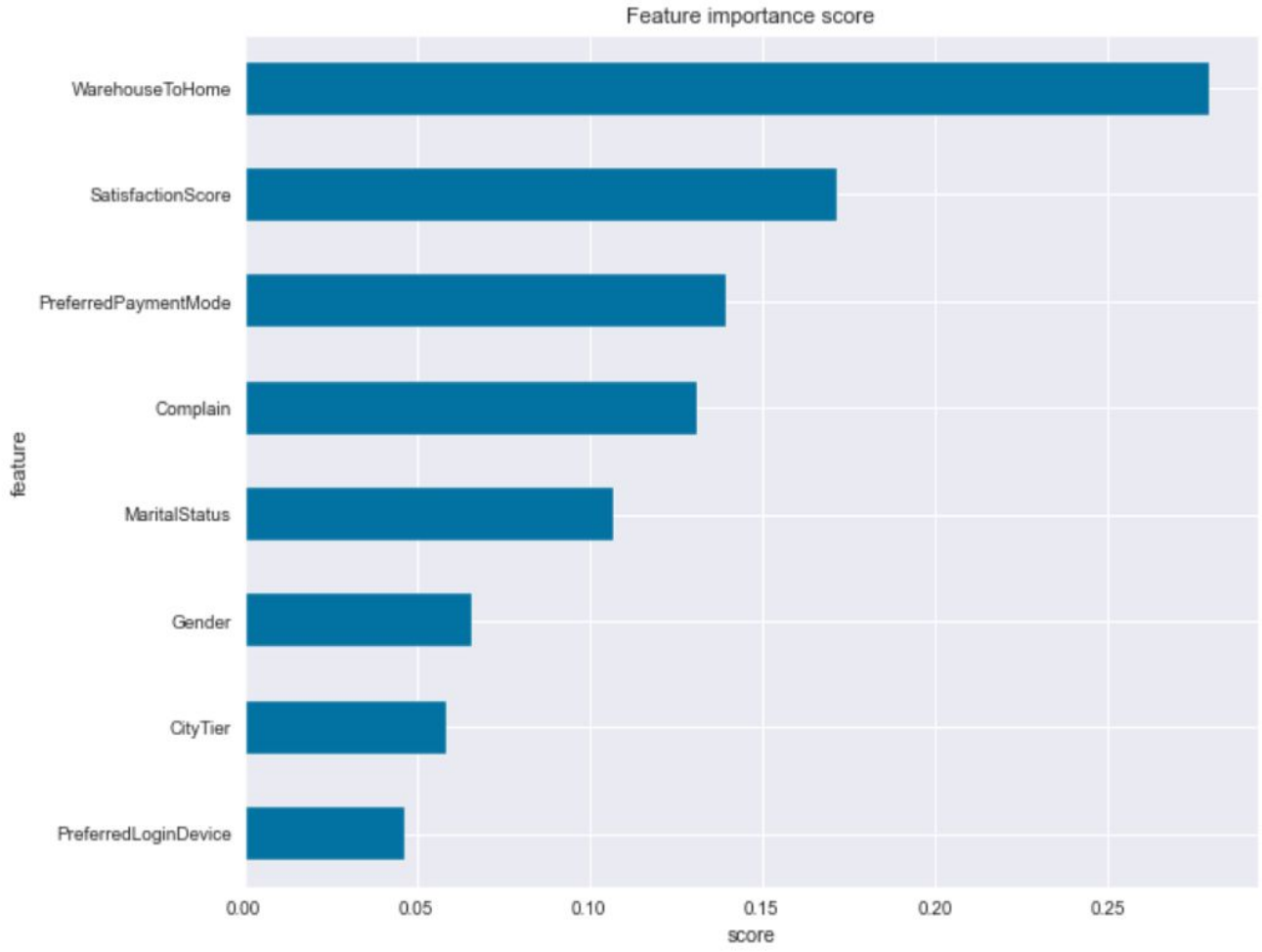
True Negatives(TN) = 150

False Positives(FP) = 30

False Negatives(FN) = 48

04 Model Evaluation (ROC & Feature Importance)





We Care About
Your Future

05 Summary and Recomendation

SUMMARY

1. Persentase pelanggan yang memilih churn sebanyak 18,4% dari total 4588 customer
2. Ada beberapa keadaan customer yang menyebabkan terjadinya churn seperti, tingkatan CityTier customer, Login Device, WarehouseToHome, PaymentMode, Complain, Gender, Marital Status dan SatisfactionScore
3. Pada data CityTier semakin tinggi tingkatan kota semakin tinggi pula potensi untuk churn. Namun khusus untuk kasus ini tingkatan kota 3 lebih rendah dikarenakan populasi di city tier 3 jauh lebih banyak ketimbang city tier 2
4. Customer yang login menggunakan computer memiliki tingkat churn yang tinggi
5. Jarak dari rumah customer menjadi salah satu penyebab terjadinya churn, semakin jauh jarak rumah customer semakin tinggi persentasi untuk churn. Bisa dikarenakan oleh ongkos kirim makin mahal dengan semakin jauh lokasi tempat tinggal dari warehouse
6. PaymentMode menjadi salah satu faktor penyebab churn, customer yang menggunakan cod lebih berpotensi untuk churn
7. Rating pada SatisfactionScore tidak mempengaruhi tingkat churn, karena Data SatisfactionScore dipengaruhi oleh beberapa faktor yang ada.
8. Model Machine learning yang digunakan untuk memprediksi Churn adalah ExtraTreesClassifier

RECOMENDATION

1. memberikan voucher gratis ongkir kepada customer
2. menambahkan diskon kepada customer yang metode bayar nya menggunakan cod
3. meningkatkan kualitas pelayanan untuk customer
4. menanggapi complain dengan cepat dan melakukan perbaikan

We Care About
Your Future

Thanks For Your Attention.

Follow our social media on :

 **@data_bangalore**

 **Data Bangalore**

 **Data Bangalore Id**