We Care About
Your Future

E-Commerce Customer Churn Prediction

Final Project

Oleh : Bayuzen Ahmad, Insan Cahya Setia, Mus'ab

We Care About
Your Future

Business Understanding

Churn rate adalah rasio pelanggan yang berhenti berlangganan dengan perusahaan dalam periode waktu tertentu. Salah satu mekanisme terbaik untuk mempertahankan pelanggan saat ini adalah mengidentifikasi potensi churn dan merespon dengan cepat untuk mencegahnya. Teknik data mining dapat diterapkan untuk menganalisis perilaku pelanggan dan untuk memprediksi pengurangan pelanggan potensial sehingga strategi pemasaran khusus dapat diadopsi untuk mempertahankannya.

Business Question

1. Berapa persen customer yang memilih untuk churn ?
2. Apakah jenis kelamin berpengaruh terhadap kemungkinan customer untuk churn?
3. Apakah ada pengaruh antara tenure dengan churn rate?
4. Apakah jarak rumah customer dengan warehouse memiliki pengaruh terhadap churn?
5. Apakah customer yang komplain cenderung akan memilih untuk churn ?
6. Bagaimana churn rate berdasarkan kategori produk yang dibeli customer ?
7. Bagaiman churn rate berdasarkan metode pembayaran yang digunakan?

Goal

1. Mengidentifikasi penyebab kemungkinan customer memilih churn.
2. Membangun model machine learning untuk mendeteksi customer yang akan churn.

We Care About
Your Future

Data Preprocessing

We Care About
Your Future

Data Preprocessing

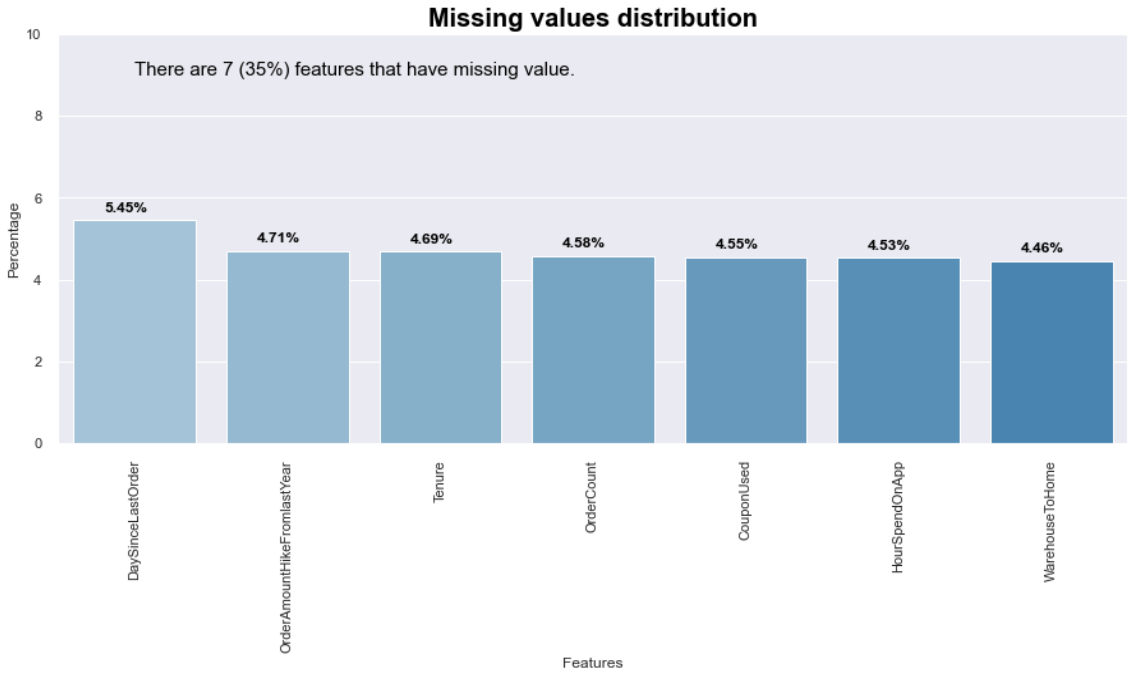
Variable	Description
CustomerID	Unique customer ID
Churn	Churn Flag
Tenure	Tenure of customer in organization
PreferredLoginDevice	Preferred login device of customer
CityTier	City tier
WarehouseToHome	Distance in between warehouse to home of customer
PreferredPaymentMode	Preferred payment method of customer
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

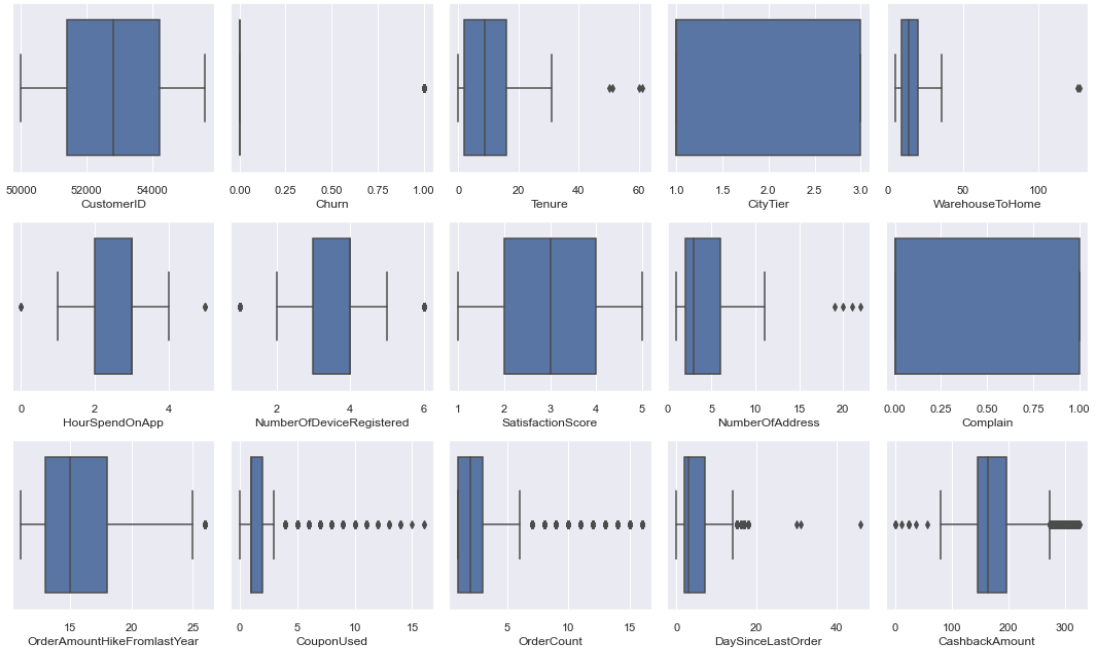
5630 rows × 20 columns

We Care About
Your Future

Data Preprocessing

Handling Missing Values





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.

Data Preprocessing

Handling Inconsistent Data

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

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

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

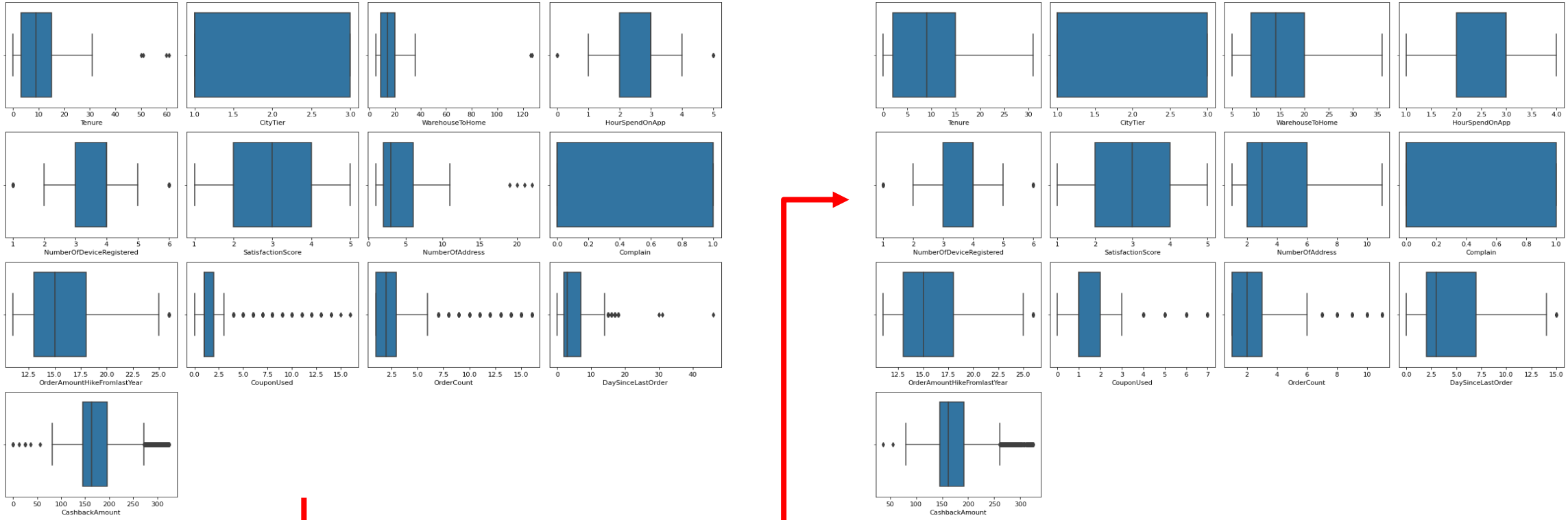
Karena terdapat inconsistent data pada beberapa feature yang seharusnya merujuk pada kategori yang sama, maka lakukan perubahan pada kategori tersebut.

```
...
    PreferredPaymentMode : ['Debit Card' 'UPI' 'Credit Card' 'Cash on Delivery' 'E wallet']

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

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

We Care About
Your Future

Data Preprocessing

Handling Outliers : Metode Z-Score



Pada kasus ini, untuk menangani data outliers, kita menggunakan metode Z-Score. Setelah data outliers dihilangkan, dimensi data berkurang sebanyak sebanyak 280 baris, dari 5630 baris menjadi 5350 baris sehingga dataset masih mengandung informasi sebesar 95%. Metode Z-Score cocok digunakan karena data outliers yang dihilangkan tidak terlalu banyak.

We Care About
Your Future

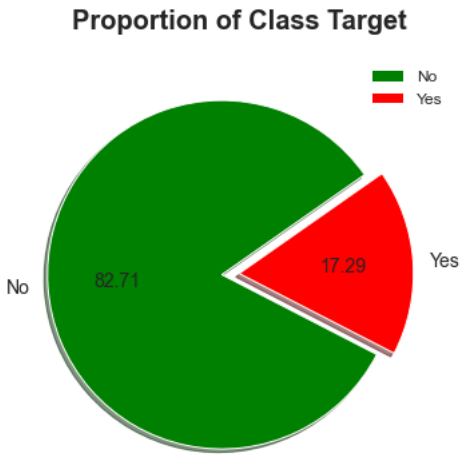
Data Preprocessing

Feature Engineering

- Remove Feature : CustomerID
- Add new features : avg_cashback_per_order, distance, tenure_category
- Feature Encoding : Ordinal Encoding and One-Hot Encoding
- Feature Scaling : Robust Scaling

Resampling Dataset

- Separate train and test set with 80% train set and 20% test set



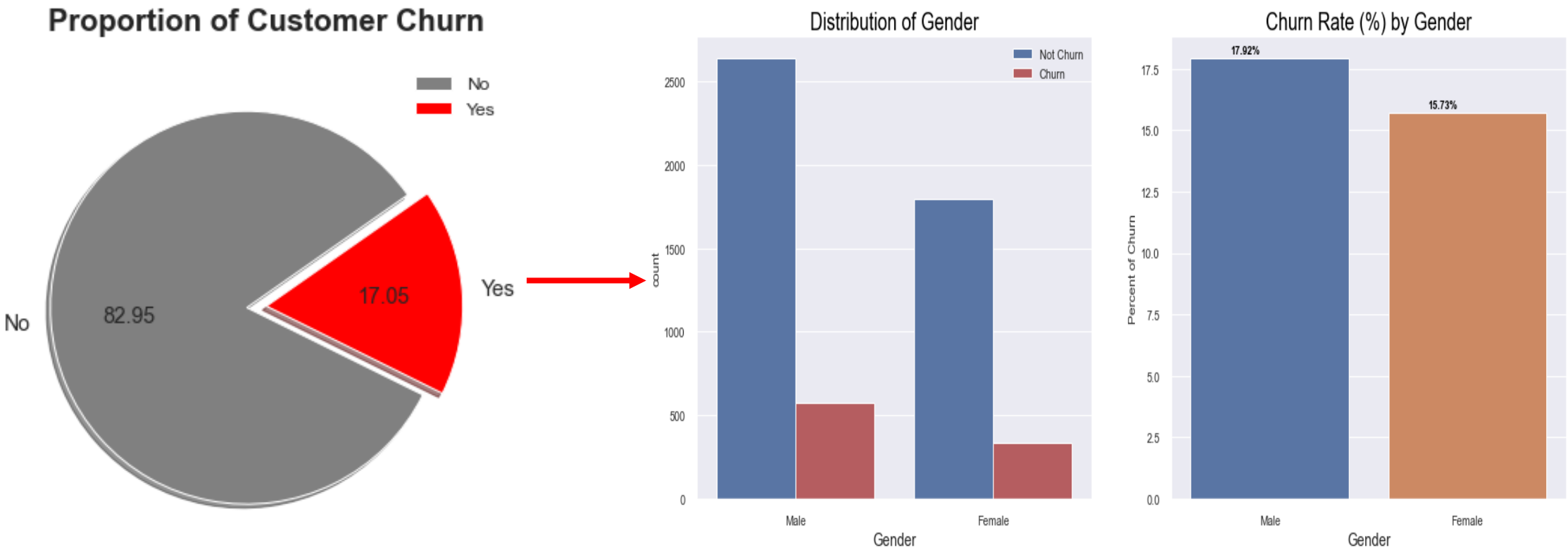
- Target imbalanced, Oversampling using SMOTE to make target to be balanced

We Care About
Your Future

Exploratory Data Analysis

We Care About
Your Future

Exploratory Data Analysis

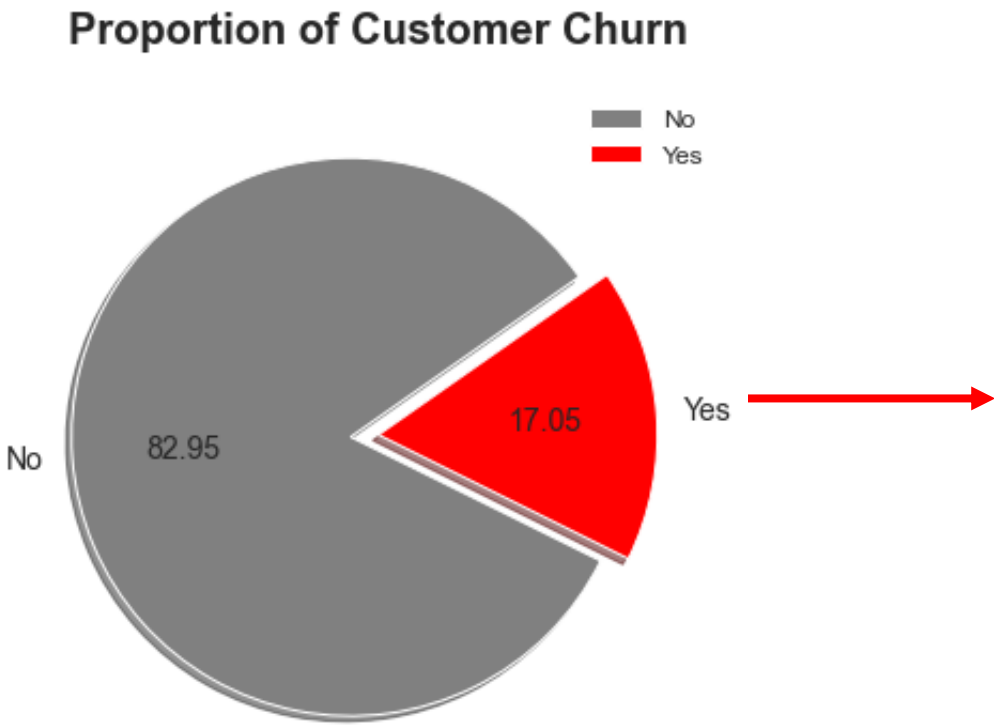


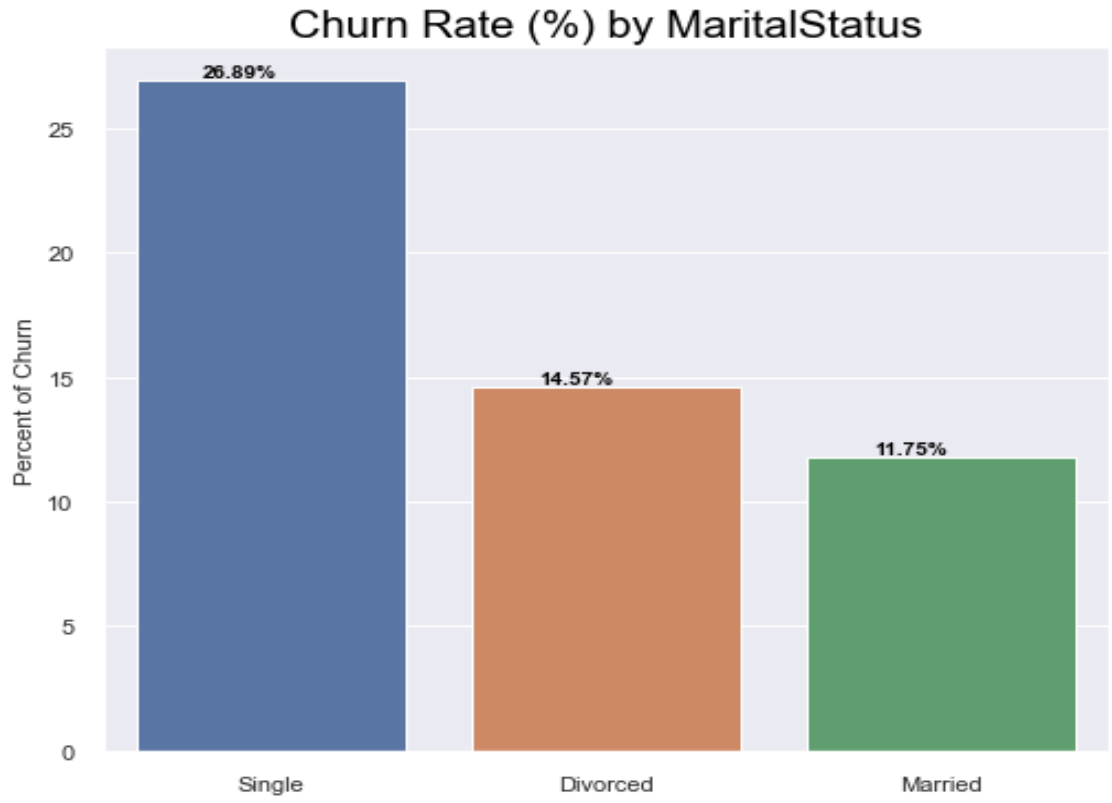
INSIGHT

1. Sekitar 17.05% atau sekitar 912 customer dari total 5350 customer yang tercatat memilih untuk churn.
2. Customer berjenis kelamin laki-laki menjadi customer dengan churn rate paling tinggi, dengan persentase sekitar 17.92% 3.
3. Customer berjenis kelamin perempuan memiliki selisih sekitar 2.19% lebih kecil dari customer laki-laki dengan persentase churn rate 15.73%.

We Care About
Your Future

Exploratory Data Analysis



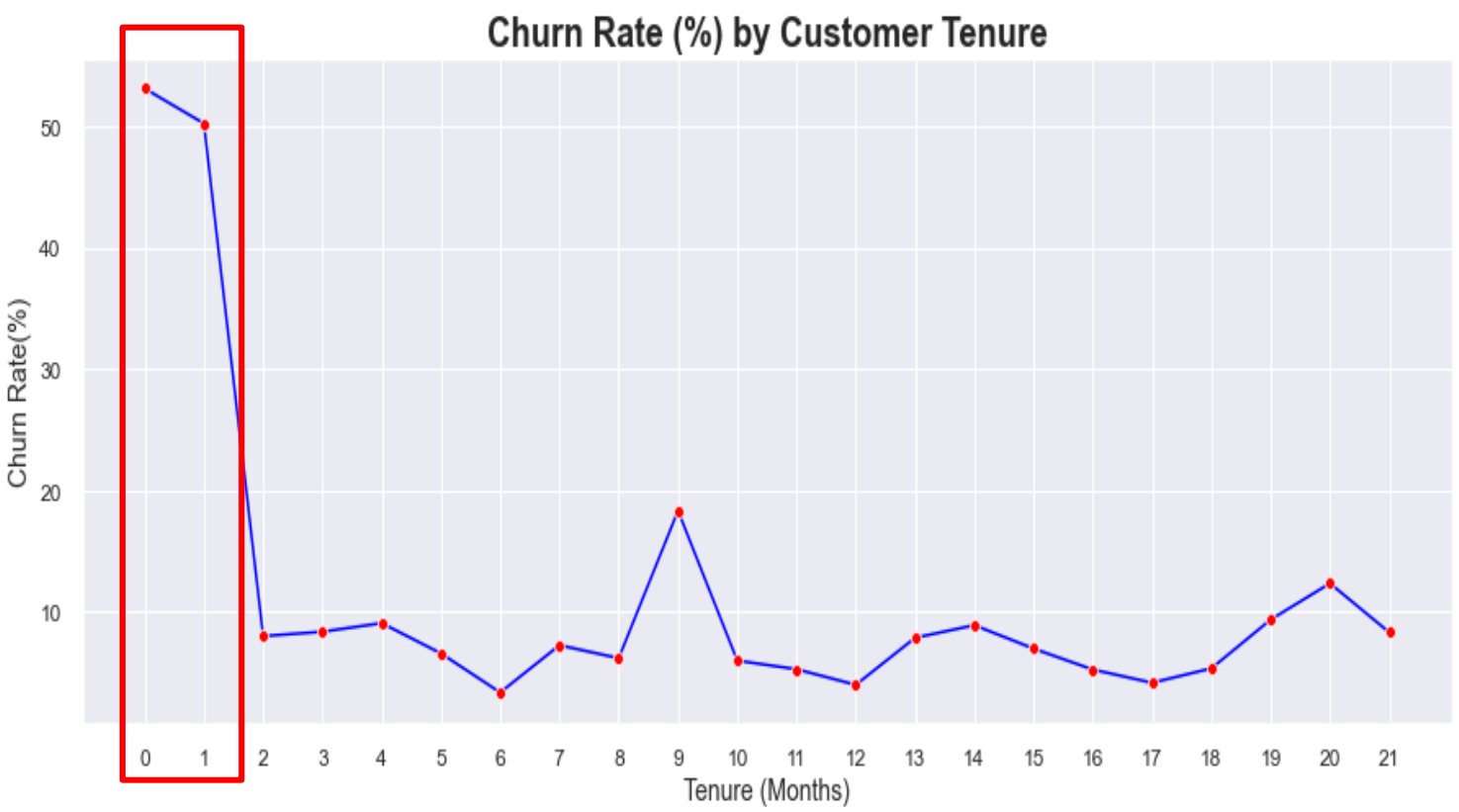


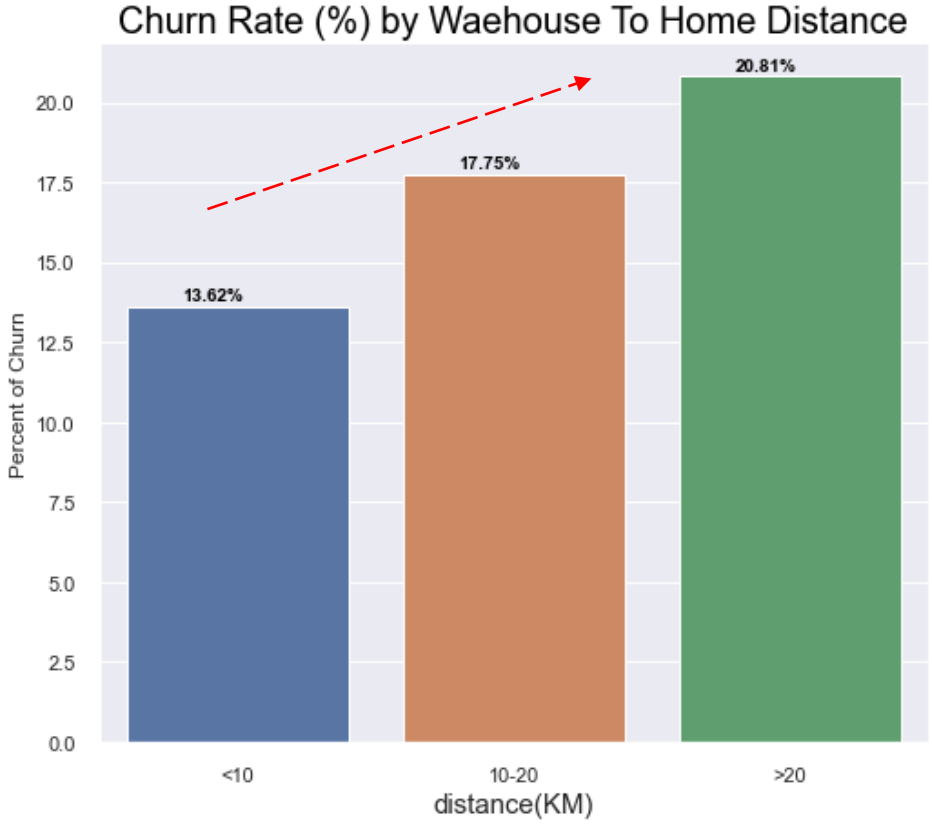
INSIGHT

Customer yang masih lajang ternyata memiliki persentase churn rate paling tinggi dari customer yang sudah menikah dan cerai, dengan persentase sekitar 26.89%

We Care About
Your Future

Exploratory Data Analysis



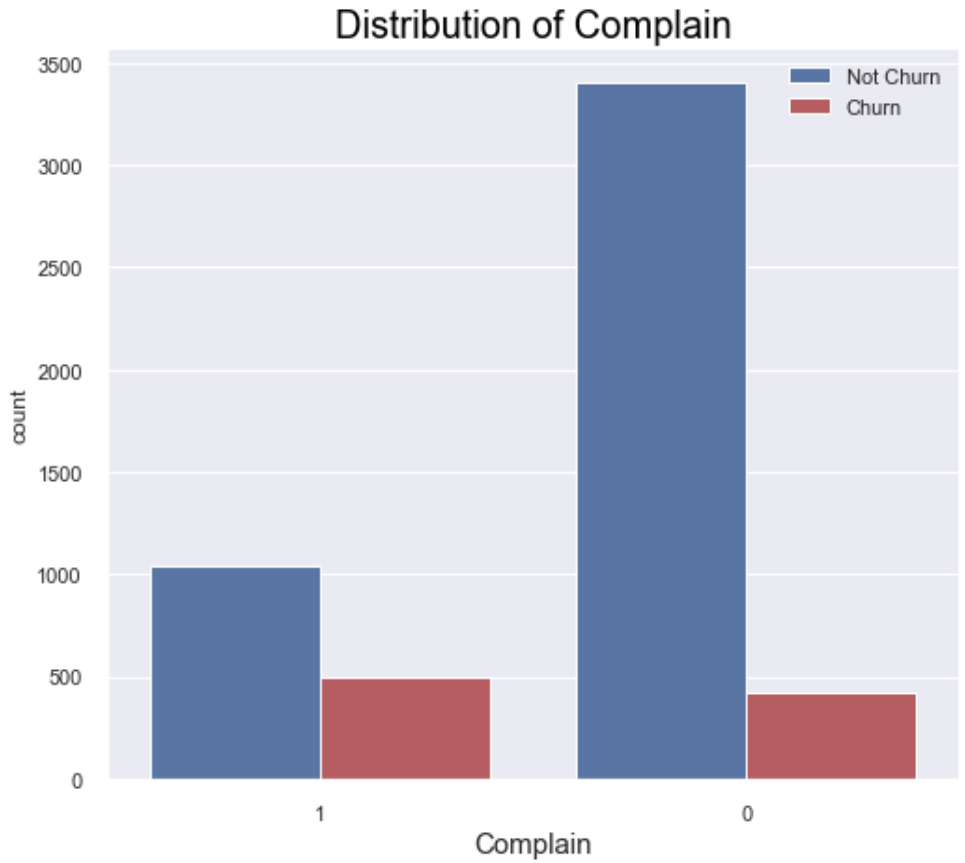


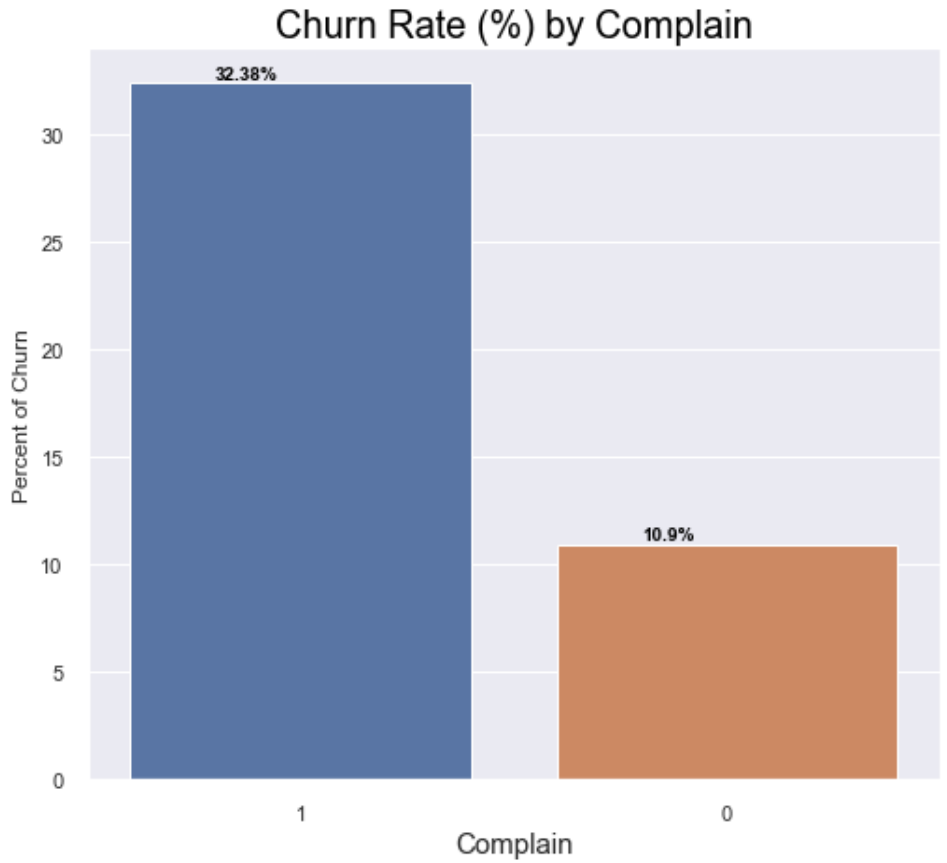
INSIGHT

1. Jika kita perhatikan, customer dengan tenure kurang dari 2 bulan cenderung memiliki churn rate tinggi, yaitu lebih dari 50%. Ketika tenure meningkat, maka churn rate pun cenderung menurun dengan persentase rata-rata kurang dari 10%. Jika dapat mempertahakan customer dengan waktu yang lama mungkin akan menyebabkan penurunan pada churn rate.
2. Churn rate rate meningkat ketika jarak antara rumah customer dengan gudang (warehouse) semakin jauh.

We Care About
Your Future

Exploratory Data Analysis



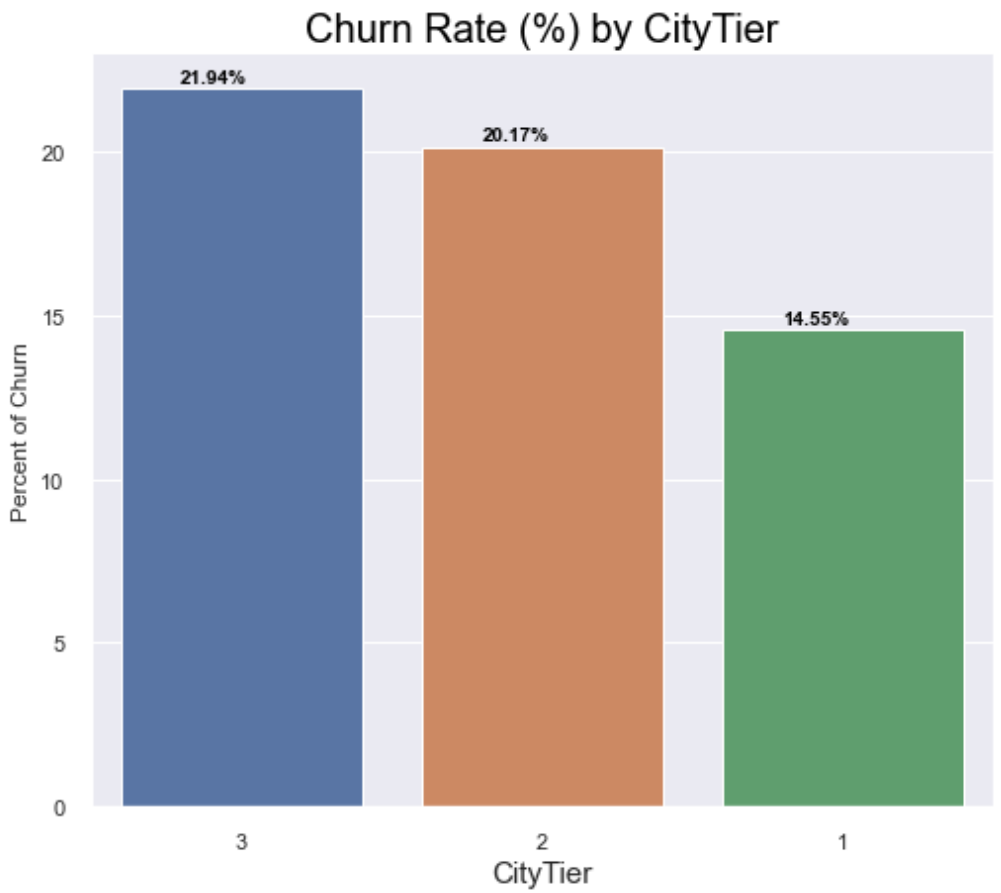


INSIGHT

1. Jika kita lihat, kebanyakan customer tidak memiliki keluhan saat mereka melakukan transaksi.
2. Customer yang memiliki keluhan cenderung memiliki churn rate lebih tinggi daripada customer yang tidak memiliki keluhan, yaitu sekitar 32.38%. Hal ini cukup masuk akal, karena ketika customer memiliki keluhan pada saat melakukan transaksi maka kemungkinan mereka untuk memilih churn pun semakin tinggi.

We Care About
Your Future

Exploratory Data Analysis

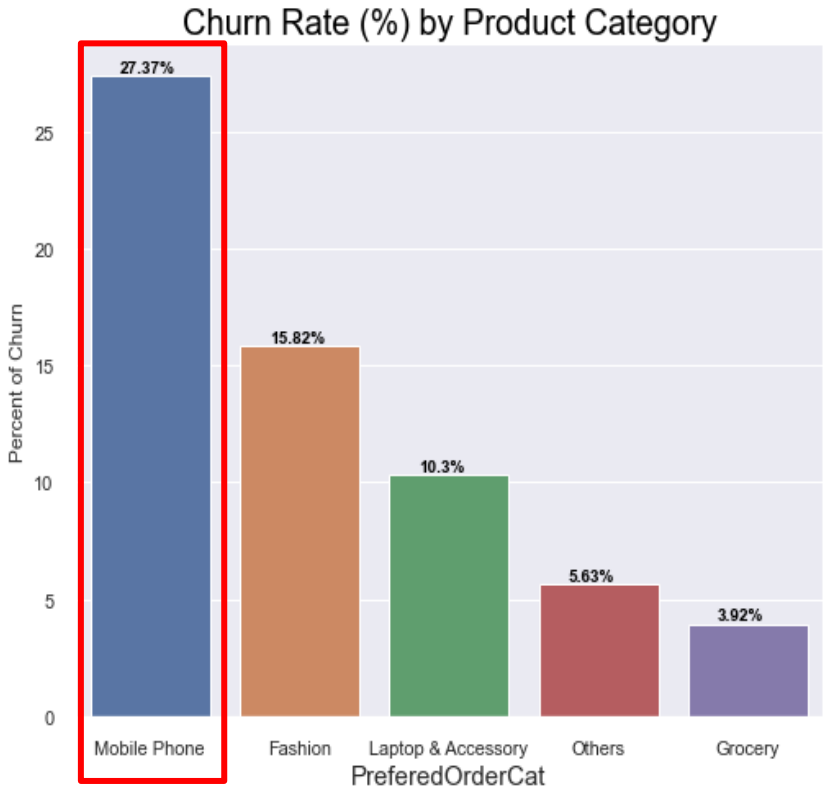


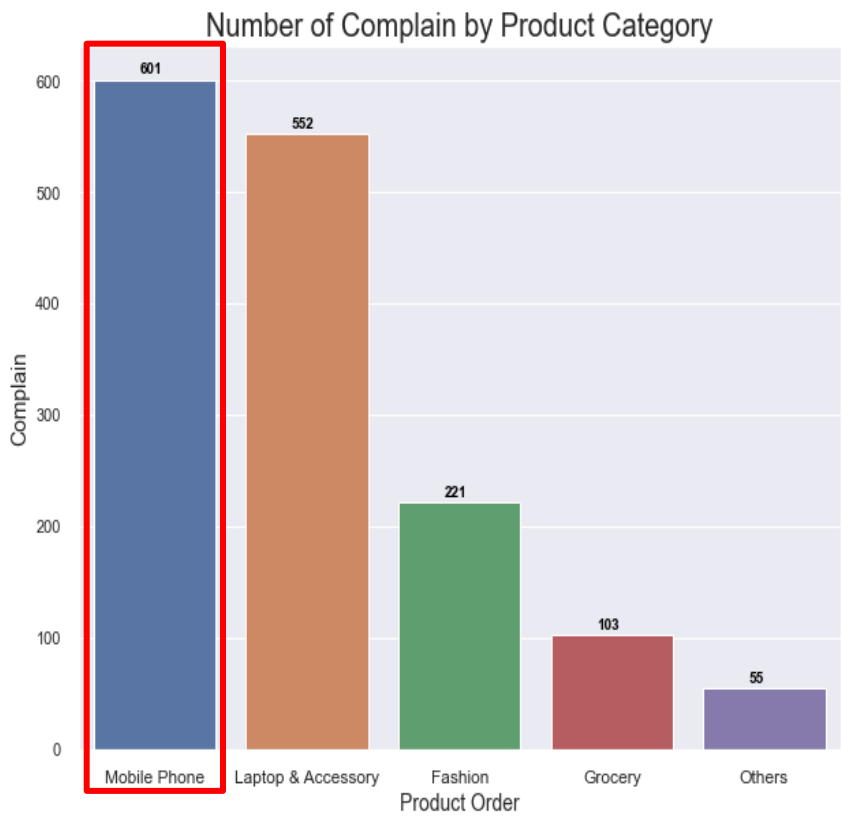
INSIGHT

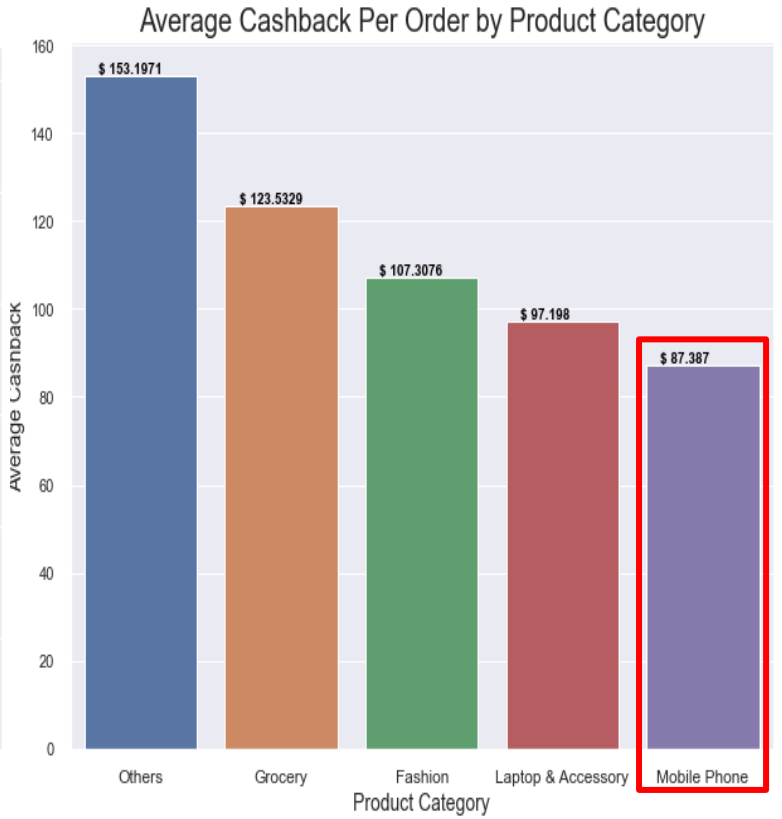
1. Customer dari kota tier 3 menjadi customer yang mengalami churn rate paling tinggi, yaitu sekitar 21.94%.
2. Meskipun jumlah customer dari kota dengan Tier 1 paling banyak, tetapi memiliki churn rate yang paling rendah, yaitu sekitar 14.55%.

We Care About
Your Future

Exploratory Data Analysis





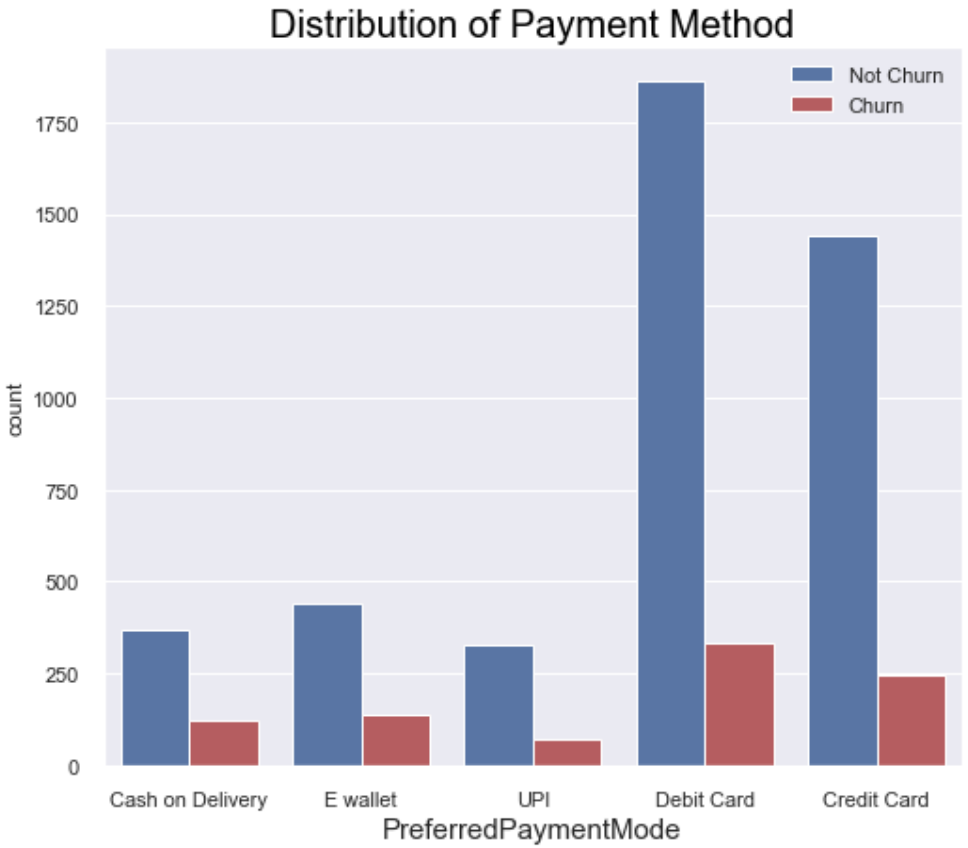


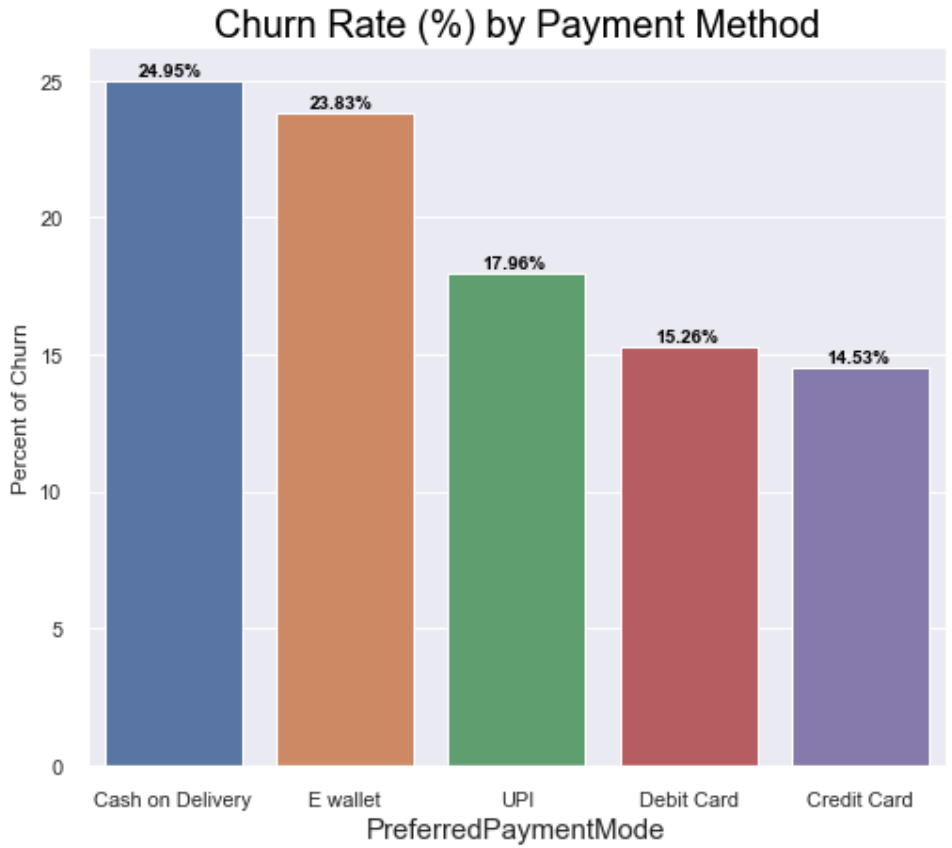
INSIGHT

1. Mobile Phone merupakan kategori produk dengan churn rate paling tinggi dari semua kategori produk, yaitu lebih dari 27%
2. Jika kita perhatikan, barang dengan komplain terbanyak yaitu termasuk barang elektronik. Bisa jadi customer merasa kecewa dengan barang yang diterima, penyebabnya bisa jadi karena tidak sesuai pesanan, atau mengalami kerusakan saat diterima oleh customer.
3. Jika kita lihat antara rata rata cashback per order dan total komplain per kategori produk, kategori produk Mobile Phone memberikan rata rata cashback yang rendah dan juga memiliki total komplain yang tinggi, sehingga hal ini mungkin saja penyebab terjadinya churn. Jika customer mendapatkan e-commerce dengan harga mobile phone yang lebih murah dan memberikan tawaran cashback yang lebih tinggi maka kemungkinan customer akan beralih ke e-commerce lain semakin tinggi.

We Care About
Your Future

Exploratory Data Analysis





INSIGHT

1. Mayoritas customer menggunakan metode pembayaran Debit Card saat bertransaksi.
2. Customer yang menggunakan metode pembayaran Cash On Delivery menjadi customer yang memiliki churn rate paling tinggi dari metode pembayaran yang lainnya, yaitu dengan persentase sekitar 24.95%.

We Care About
Your Future

Modelling and Evaluation

We Care About
Your Future

Modelling and Evaluation

Choosing the Best Model

1. Membandingkan beberapa model machine learning yaitu KNN, Logistic Regresion, SVM, Naïve Bayer, MLP, Decision Tree, Random Forest, ExtraTrees, AdaBoost, GradientBoosting, Bagging, XGBoost.
2. Metode evaluasi menggunakan cross validation dan Recursive Feature Elimination untuk feature selection.
3. Metrik evaluasi yang digunakan adalah accuracy, precision, recall dan f-1 score.
4. Selain nilai akurasi yang digunakan, kita juga ingin model kita banyak False Positive daripada False Negative. Maka nilai recall pun akan kita perhitungkan.

	Model	fit_time	score_time	test_accuracy	test_precision	test_recall	test_f1
0	XGBClassifier	2.514655	0.025586	0.994393	0.983622	0.983547	0.983546
1	ExtraTreesClassifier	1.206610	0.148457	0.993458	1.000000	0.961635	0.980344
2	RandomForestClassifier	1.376692	0.170867	0.989533	0.991036	0.947337	0.968458
3	DecisionTreeClassifier	0.086550	0.022365	0.985607	0.951505	0.964907	0.958101
4	BaggingClassifier	0.556066	0.027426	0.980374	0.963785	0.919972	0.941126
5	MLPClassifier	11.257185	0.022924	0.968224	0.936284	0.873945	0.903460
6	KNeighborsClassifier	0.009918	0.518106	0.922617	0.893178	0.620621	0.732242
7	GradientBoostingClassifier	2.162448	0.020611	0.921121	0.845352	0.658998	0.740171
8	SVC	1.362156	0.869904	0.913832	0.890933	0.563670	0.689508
9	AdaBoostClassifier	0.814248	0.065768	0.896636	0.750368	0.589996	0.659951
10	LogisticRegression	0.177303	0.019189	0.895327	0.780061	0.538450	0.636524
11	GaussianNB	0.025094	0.046747	0.677383	0.319730	0.787306	0.454544

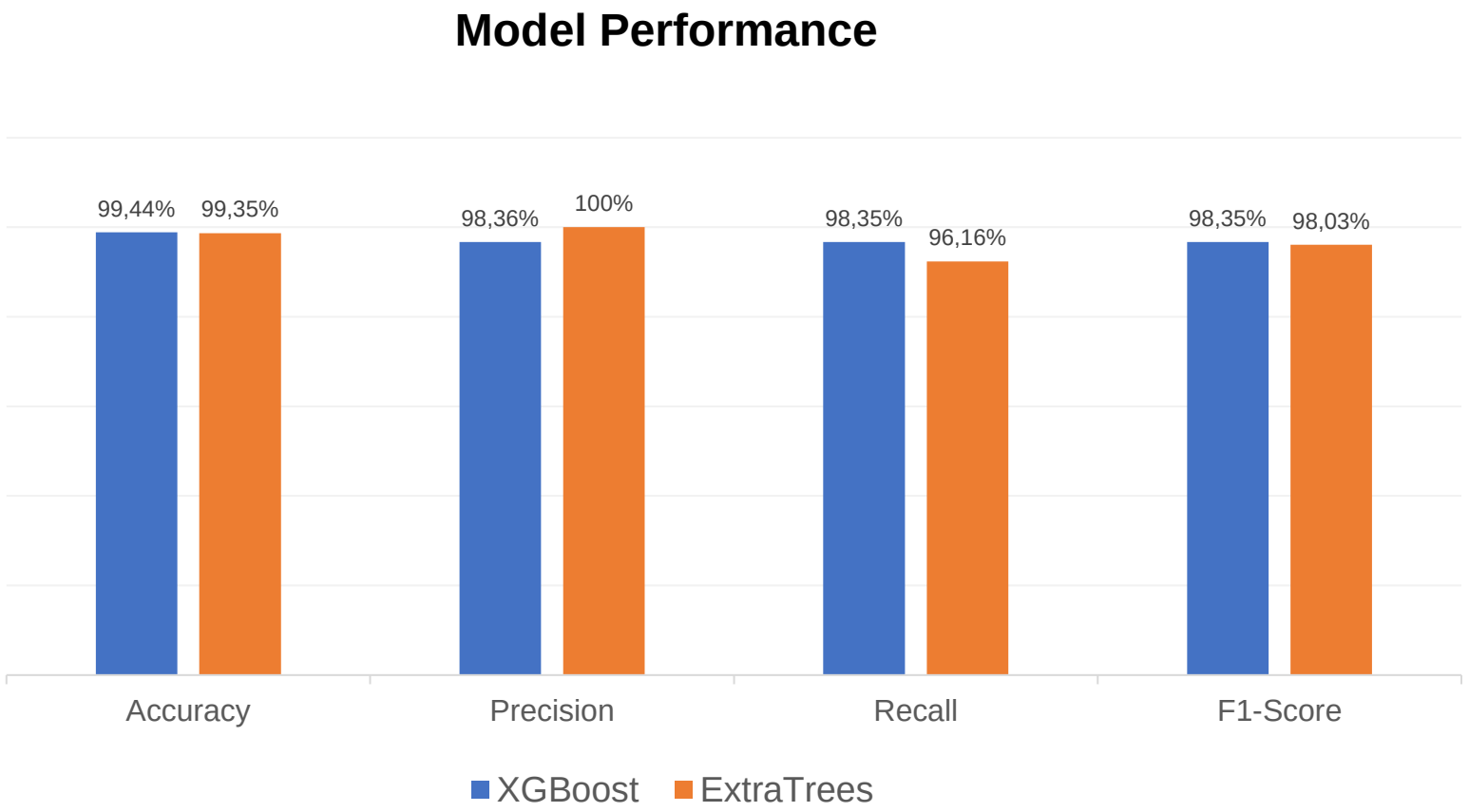
Best Model

We Care About
Your Future

Modelling and Evaluation

Cross Validation

Dari hasil evaluasi menggunakan cross validation, di dapatkan dua model terbaik, yaitu XGBoost dan ExtraTrees.



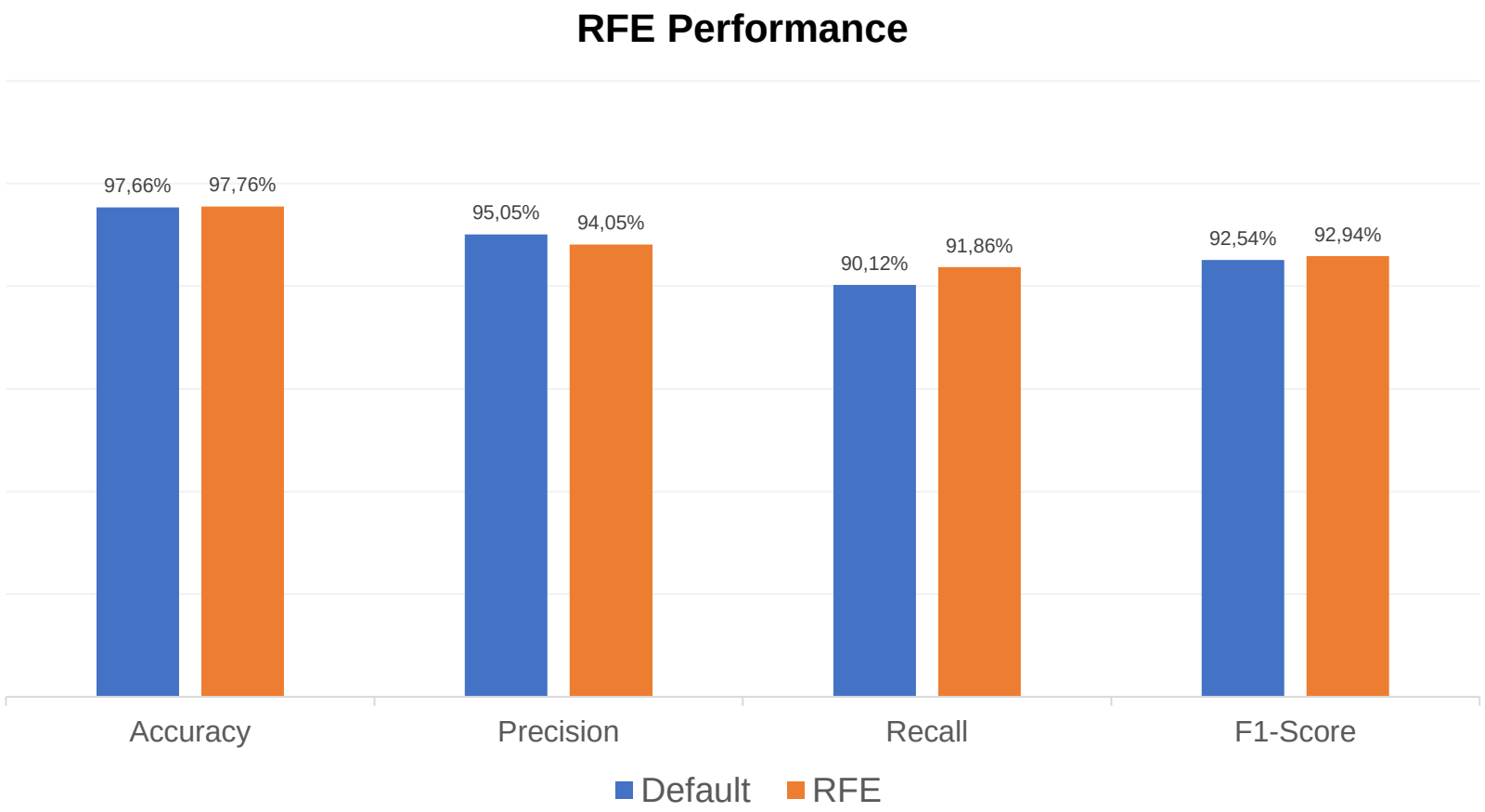
Karena XGBoost merupakan model dengan performa yang paling baik, maka kita memilih XGBoost untuk tahap modelling selanjutnya.

We Care About
Your Future

Modelling and Evaluation

Recursive Feature Elimination

Optimal number of features = 25 features



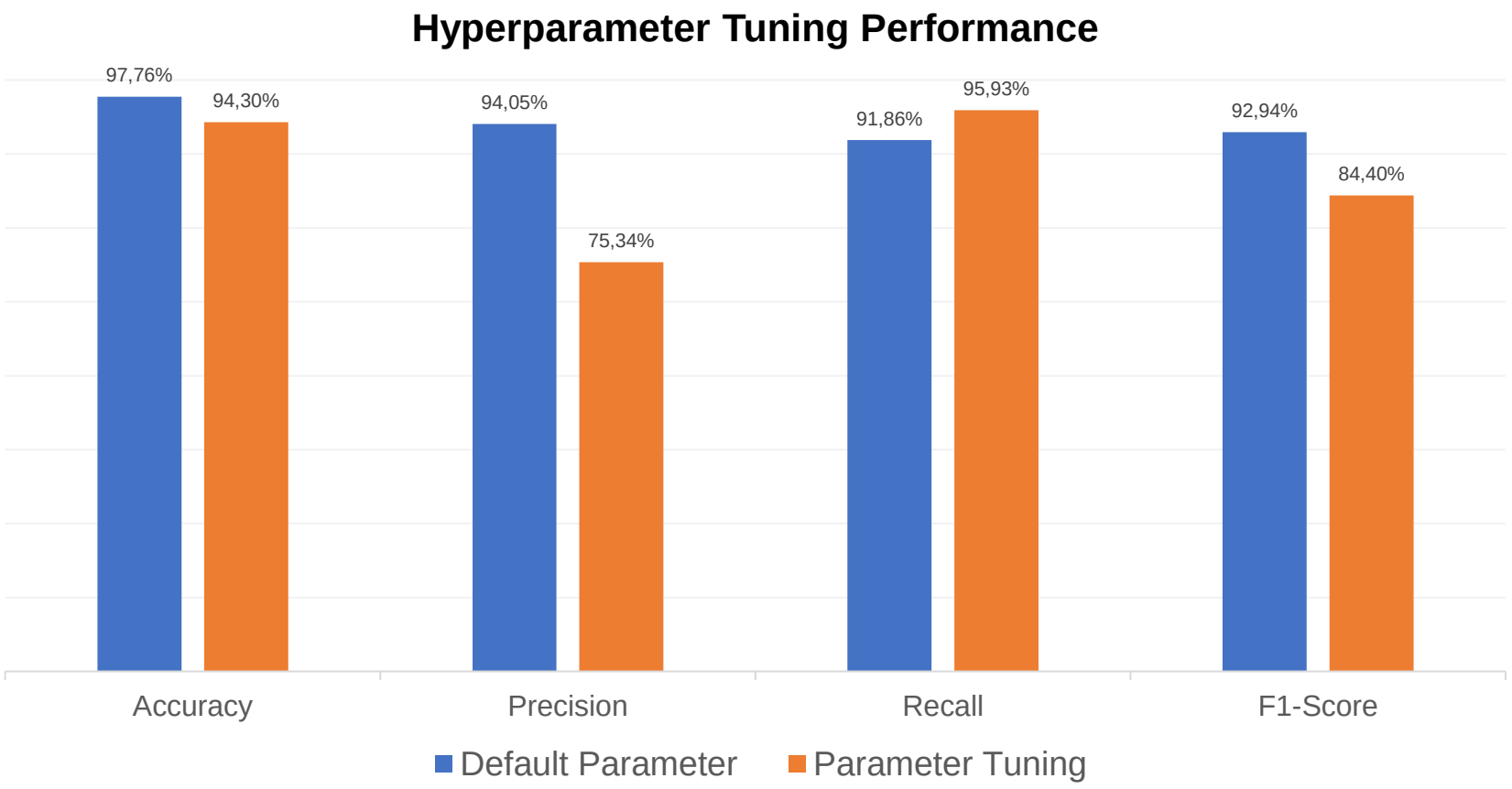
Karena hasil dari RFE memiliki performa lebih baik, maka kita gunakan features hasil dari RFE.

We Care About
Your Future

Modelling and Evaluation

Hyperparameter Tuning

Lakukan hyperparameter tuning untuk mendapatkan parameter terbaik.
Metode tuning yang digunakan yaitu Random Search CV.



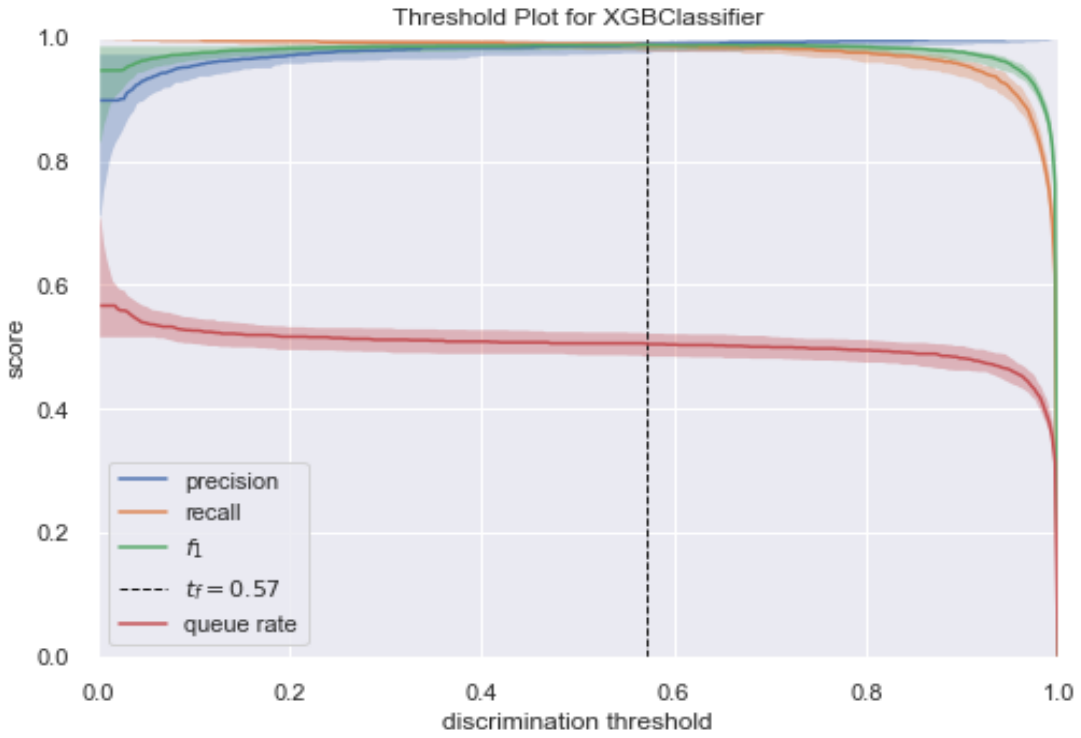
Dari hasil parameter tuning, ternyata tuning parameter tidak cukup baik untuk membuat model lebih baik. Maka kita tetap akan menggunakan parameter default.

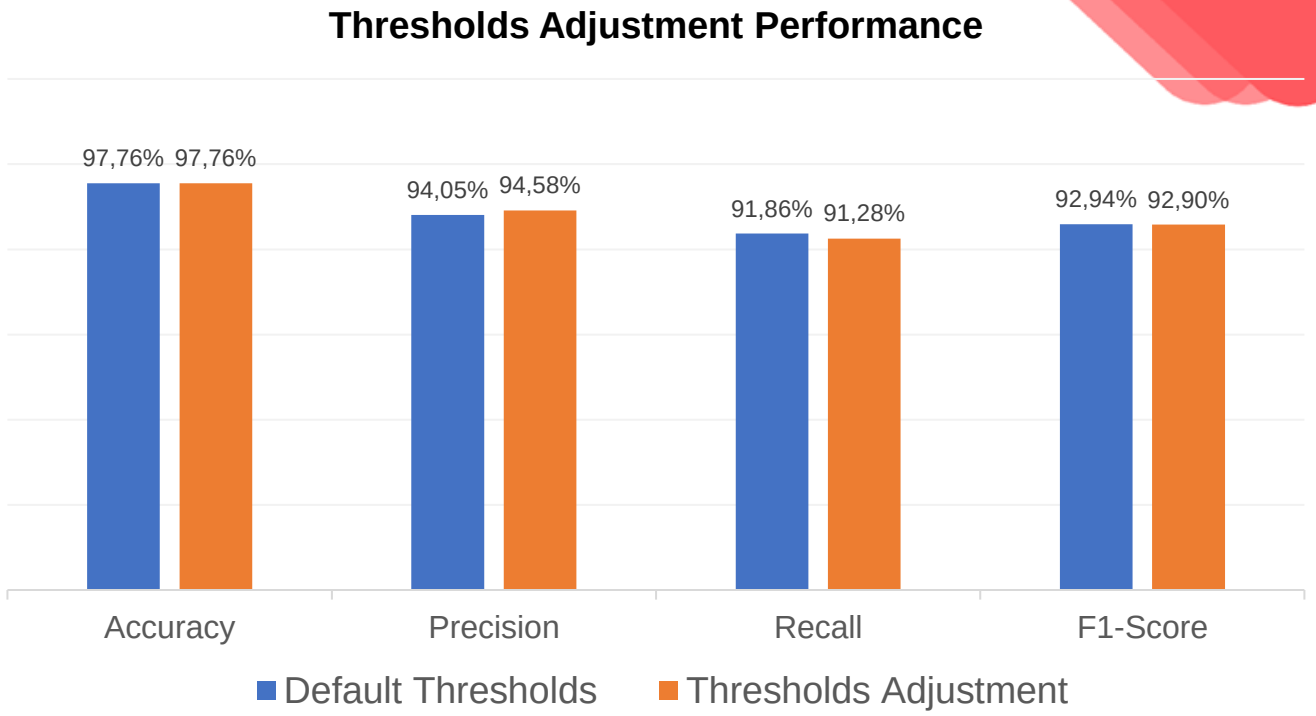
We Care About
Your Future

Modelling and Evaluation

Thresholds Adjustment

Menentukan nilai thresholds terbaik menggunakan Discrimination Thresholds. Maka didapat nilai thresholds terbaik yaitu 0.57.

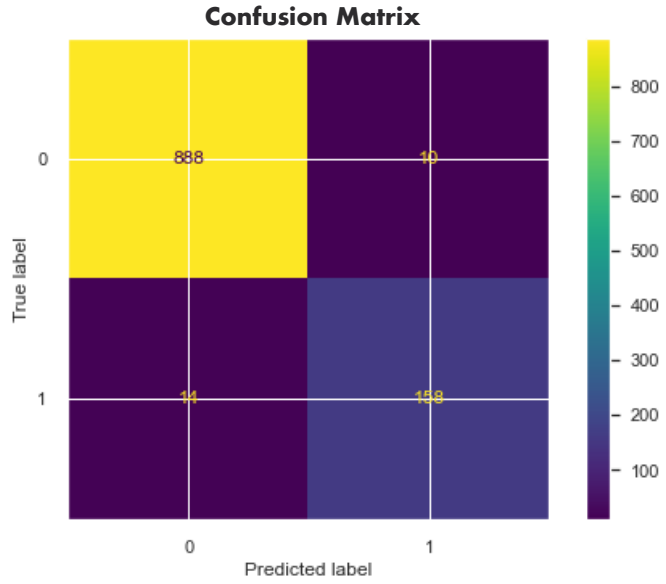


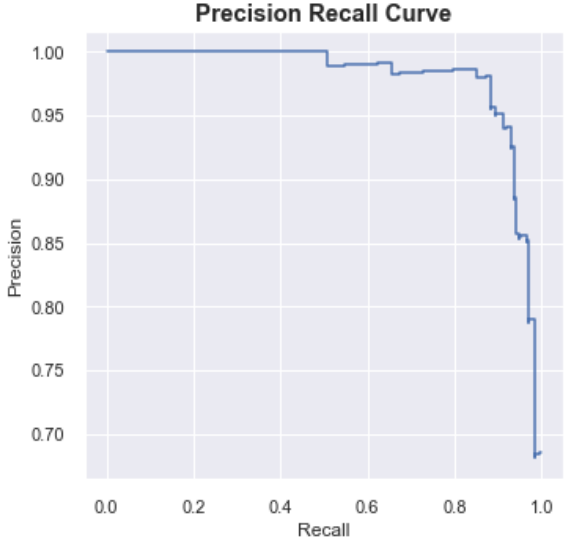


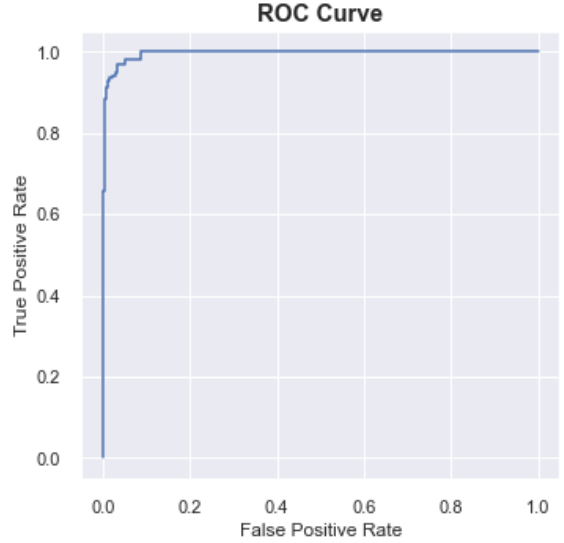
Ternyata setelah dilakukan thresholds adjustment, hasil yang didapatkan mengalami penurunan performa, maka kita tidak akan menggunakan nilai thresholds tersebut.

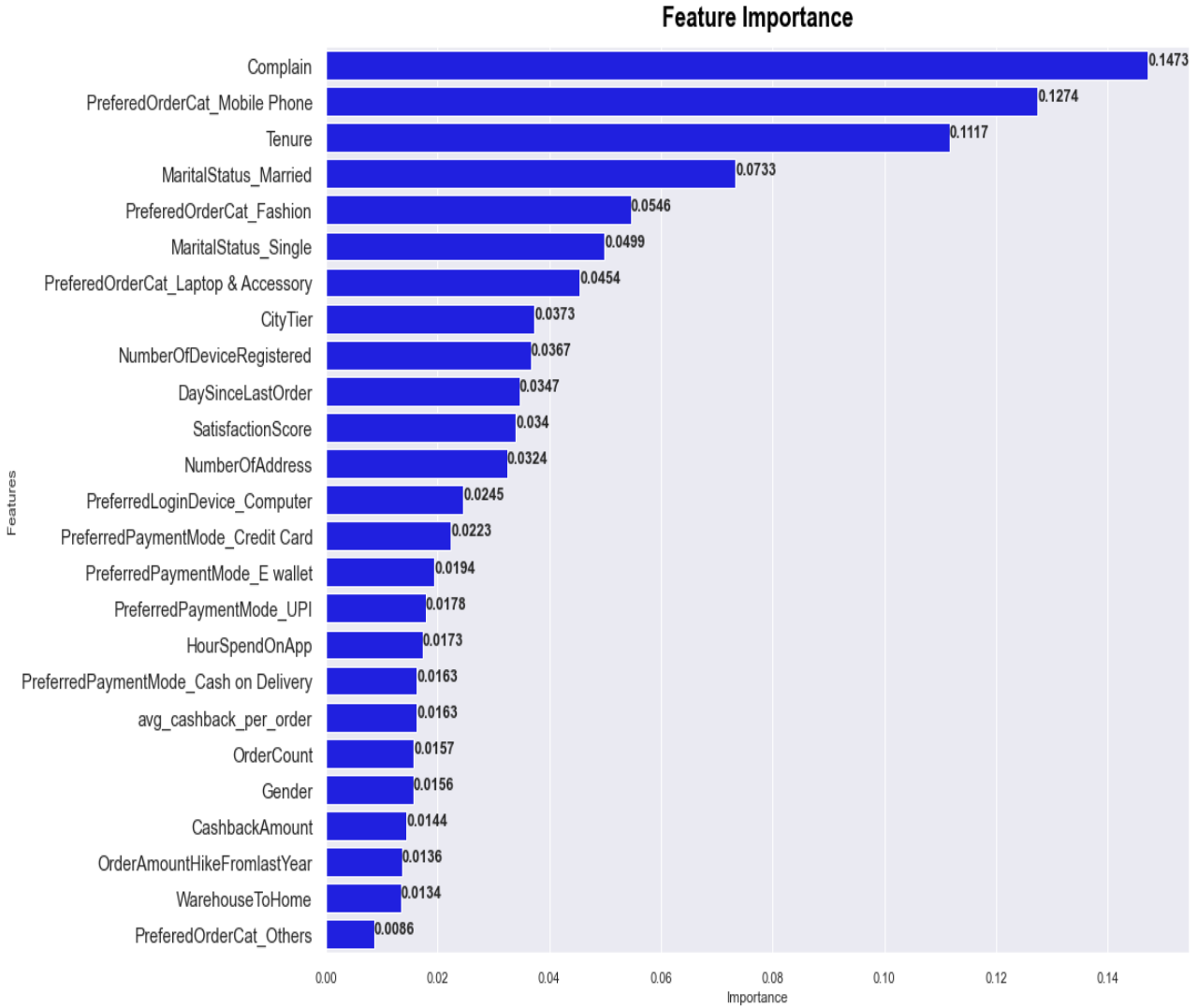
We Care About
Your Future

Modelling and Evaluation









We Care About
Your Future

Summary & Recommendation

Summary

1. Dari total 5350 customer yang tercatat, sebanyak 17.05% customer memilih untuk churn.
2. Customer dengan jenis kelamin laki-laki menjadi customer dengan churn rate tertinggi, yaitu 17.42%
3. Customer dengan tenure kurang dari 2 bulan cenderung memiliki churn rate tinggi, yaitu lebih dari 50%.
4. Churn rate semakin meningkat ketika jarak antara rumah dengan warehouse semakin jauh.
5. Customer yang memiliki komplain cenderung memiliki churn rate lebih tinggi daripada customer yang tidak memiliki komplain, yaitu sekitar 32.38%.
6. Customer yang membeli produk Mobile Phone cenderung memiliki churn rate paling tinggi dari semua kategori produk yang dibeli, yaitu lebih dari 27%.
7. Customer yang menggunakan metode pembayaran Cash On Delivery menjadi customer yang memiliki churn rate paling tinggi dari metode pembayaran yang lainnya, yaitu dengan persentase sekitar 24.95%.

Recommendation

1. Menambahkan ads untuk customer yang akan churn.
2. Untuk setiap pembelian barang, dapat voucher gratis ongkir.
3. Menawarkan membership.
4. Melakukan evaluasi review customer terhadap produk.

We Care About
Your Future

Thanks For Your Attention.

Follow our social media on :

@data_bangalore Data Bangalore Data Bangalore Id