Customer Churn Risk Score : EDA & Supervised Model

Portofolio Data Science

GitHub:
bit.ly/AmmaryfDS2

Muammar Yusuf Fakhri

Memahami Masalah dan Dataset

Customer Churn

Kehilangan pelanggan dari suatu bisnis. *Churn* dihitung dari berapa banyak pelanggan meninggalkan bisnis Anda dalam waktu tertentu. *Customer churn* penting diketahui bisnis karena merupakan ukuran kesuksesan suatu bisnis dalam mempertahankan *customer*

Tingkat *churn* yang tinggi dapat berdampak negatif pada bisnis dan juga dapat menunjukkan ketidakpuasan terhadap produk atau layanan. Bisnis akan rugi besar jika kehilangan pelanggan.

Beberapa Cara Menghentikan *Churn* :
1. Cari tahu penyebab *churn* dari customer *complaint* dan *feedback*
2. Tingkatkan *customer* engagement dengan memberikan *reward* berupa discount atau Penawaran lainnya
3. Fokus pada *customer* yang berlangganan (*Membership*)
4. Memperhatikan *tren churn rate*

Dataset

HackerEarth: How NOT to lose a customer in 10 days

Dataset ini termasuk dalam Tantangan *Machine Learning* yang diselenggarakan di Hacker Earth diposting pada tanggal 16 Maret 2021

Tujuannya Menganalisa penyebab terjadinya *Churn* dan Memprediksi tingkat risiko *churn*

Dataset yang digunakan adalah Train.csv berisi 25 *column (features), 36992 rows*

Setiap data diberi nilai prediksi yang memperkirakan status *customer churn* pada waktu tertentu, berdasarkan :

1. Informasi *Customer*
2. *Browsing Behavior*
3. *Historical Purchase*

24 Features

10 Informasi Customer

- customer_id
- Name
- age
- gender
- security_no
- region_category
- membership_category
- joining_date
- joined_through_referral
- referral_id

6 Browsing Behavior

- medium_of_operation
- internet_option
- last_visit_time
- days_since_last_login
- avg_time_spent
- avg_frequency_login_days

8 Historical Purchase

- preferred_offer_types
- avg_transaction_value
- point_in_wallet
- used_special_discount
- offer_aplicatio_preference
- past_complaint
- complaint_status
- feedback

1 Target

Churn

- churn_risk_score

1 Target → Churn

churn_risk_score

1 2 3 4 5

Nilai 1 memiliki kemungkinan pelanggan terkecil untuk melakukan churn
Nilai 5 memiliki kemungkinan pelanggan terbesar untuk melakukan churn

Nilai -1 pada feature *churn* akan di drop dan tidak akan digunakan dalam *explatory data analyst* dan *model development*.Dataset dari 36992 rows tersisa 35829 row

Tipe Data Feature

Categorycal :

1. customer_id
2. Name
3. gender
4. security_no
5. region_category
6. membership_cetgory
7. joined_through_referral
8. referral_id
9. preferred_offer_types
10. medium_of_operation
11. internet_option
12. used_special_discount
13. offer_aplication_preference
14. past_complaint
15. complaint_status
16. feedback
17. churn_risk_score

Numerical :

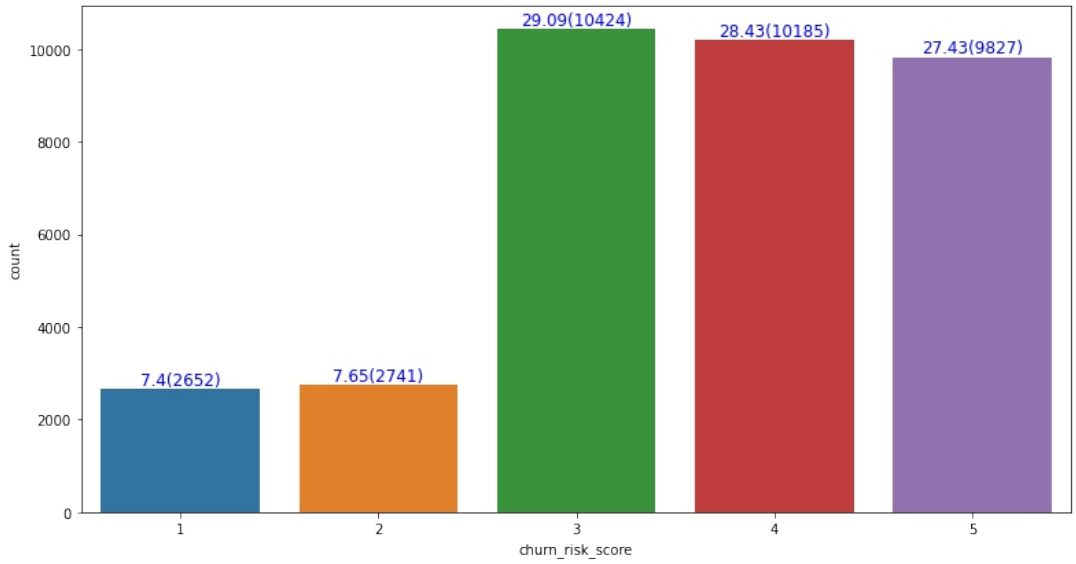
1. age
2. days_since_last_login
3. point_in_wallet
4. avg_time_spent
5. avg_transaction_value
6. avg_frequency_login

Datetime :

1. last_visit_time
2. joining_date

Explatory Data Analysis

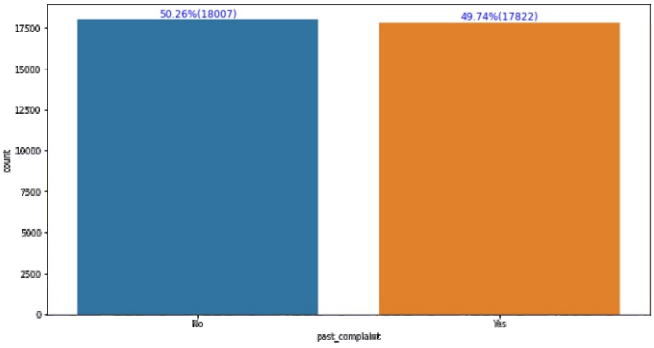
Churn Risk Score

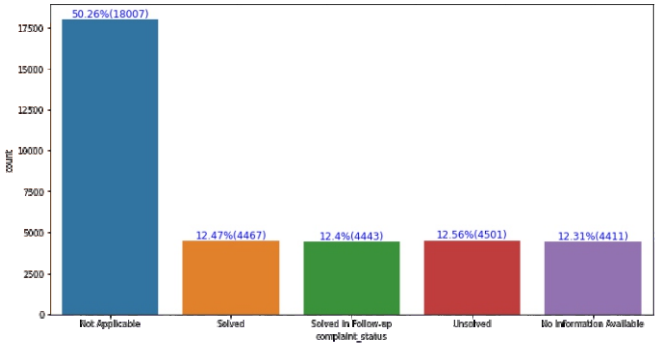


Bar chart diatas menunjukan bahwa perbandingan tingkat churn lebih banyak pada churn risk score 3, 4, dan 5 dibandingkan dengan tingkat churn1 dan 2.

Ini menunjukan bahwa banyak pelanggan yang tidak puas terhadap produk atau layanan. Ini berdampak buruk terhadap Bisnis. Apa yang membuat tingkat Churn tinggi?

Complaint

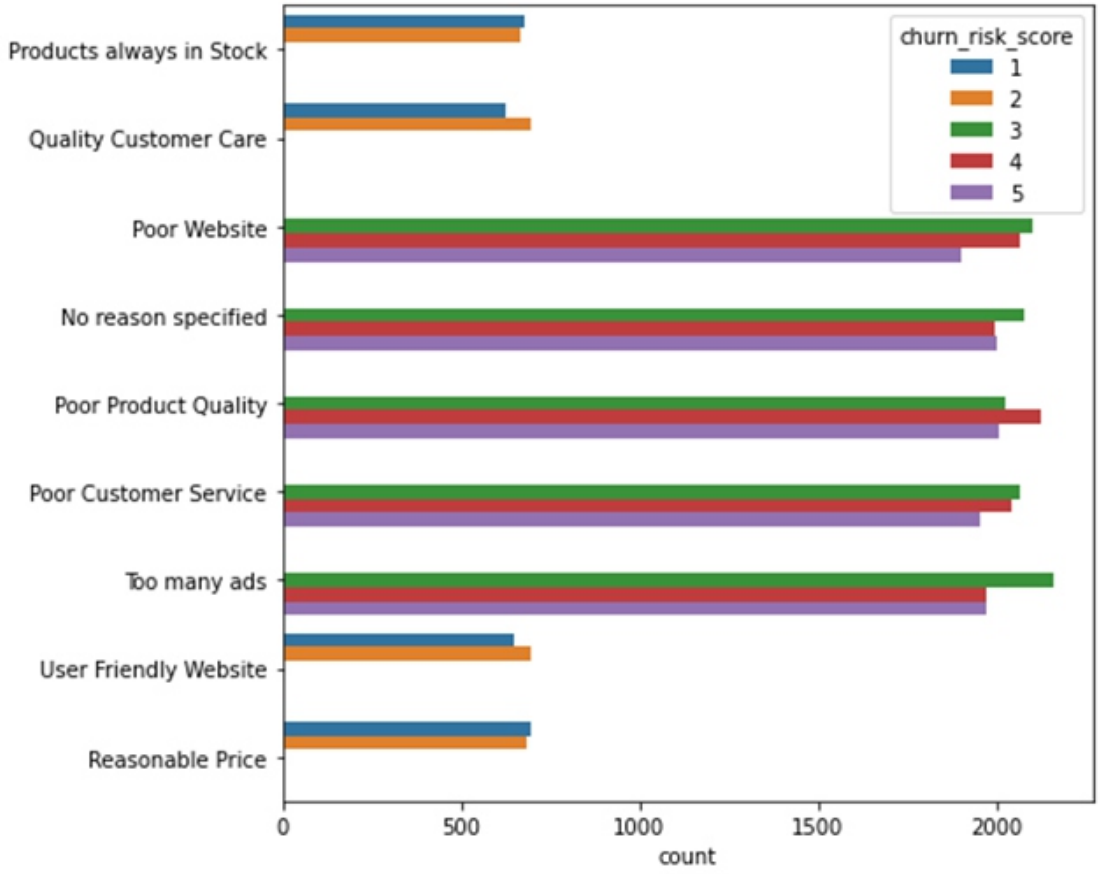




Dilihat dari fitur past_complaint, selisih antara yang mengeluh dan tidak mengeluh adalah 0,52%. Bisnis yang baik harus memiliki perbedaan perbandingan yang lebih besar, tentunya dengan proporsi pelanggan yang mengeluh lebih kecil.

Melihat lebih detail pada fitur complaint_status, masih ada complaint yang diabaikan dilihat dari proporsi *no information avaliable* sebesar 12,31%, Churn rate akan tinggi ketika ada complaint dari customer diabaikan atau tidak diperhatikan oleh bisnins

Feedback



Bar chart menunjukkan bahwa Untuk customer dengan *Churn Risk Score* 1 dan 2 memberikan feedback positif,yaitu:

1. Products always in Stock
2. Quality Customer Care
3. Use Friendly Website
4. Reasonable Price

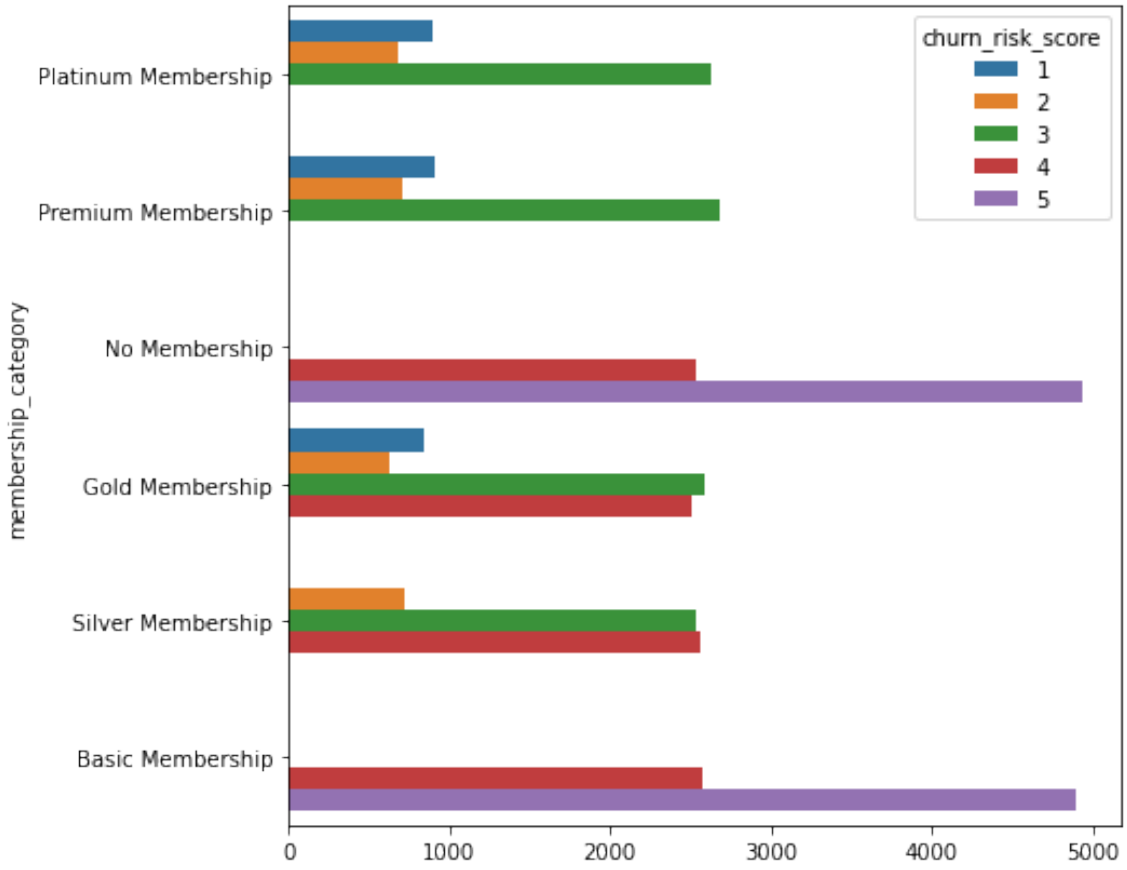
Bar chart menunjukkan bahwa Untuk customer dengan *Churn Risk Score* 3,4 dan 5 memberikan feedback negatif,yaitu:

1. Poor Website
2. Poor Product Quality
3. Poor Customer Service
4. Too many ads

Customer dengan Churn Risk Score 3, 4 dan 5 juga ada yang netral atau tidak memberikan feedback negatif maupun positif dilihat dari No Reason Specified.

Dapat disimpulkan bahwa bisnis perlu fokus memperbaiki service terhadap customer dengan churn risk score 3, 4 dan 5 karena ada yang memberikan feedback negatif.

Membership



Urutkan Membership_category dari terendah ke tertinggi :
1. No Membership
2. Basic Membership
3. Silver Membership
4. Gold Membership
5. Premium Membership
6. Platinum Membership

Bar chart menunjukkan membership category kedua tertinggi yaitu Platinum dan Premium Membership terdapat customer yang churn risk score 1, 2 dan 3. Churn risk score 3 terbanyak

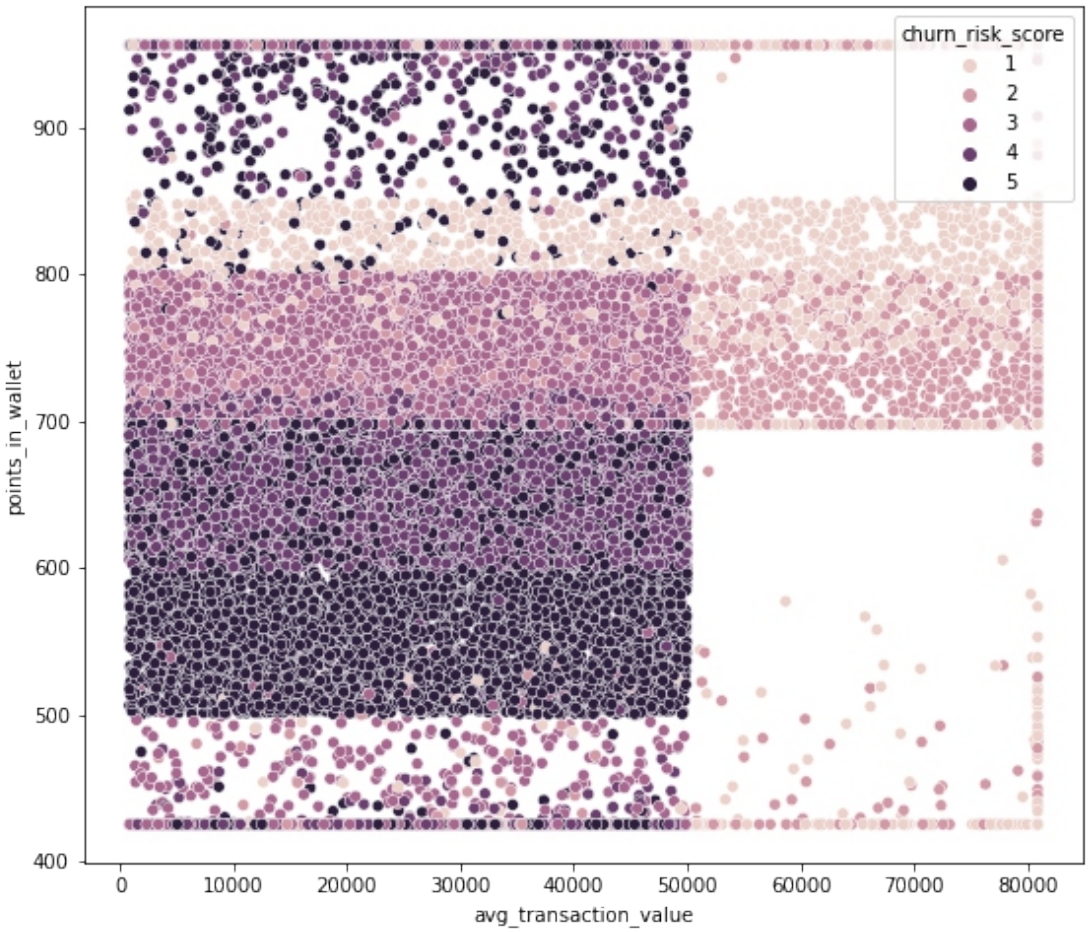
Turun ke Gold Membership, mulai ada customer dengan churn risk score 4.

Turun ke Silver Membership, sudah mulai tidak ada customer dengan churn risk score 1

Kemudian Basic Membership dan No Membership hanya ada churn risk score 4 dan 5

Ingat pada bar chart feedback bahwa churn risk score 3, 4 dan 5 adalah customer dengan feedback negatif. ini bisa menjadi evaluasi mengapa customer yang sudah menjadi membership (Basic Membership - Platinum Membership) masih memiliki feedback negatif atau complaint. Bisnis harus fokus memperbaiki pelayan terhadap mereka.

Distrbusi Customer



Scatterplot menunjukkan bahwa sebaran customer yang telah melakukan transaksi lebih dari 50000 memiliki churn risk score 1 dan 2.

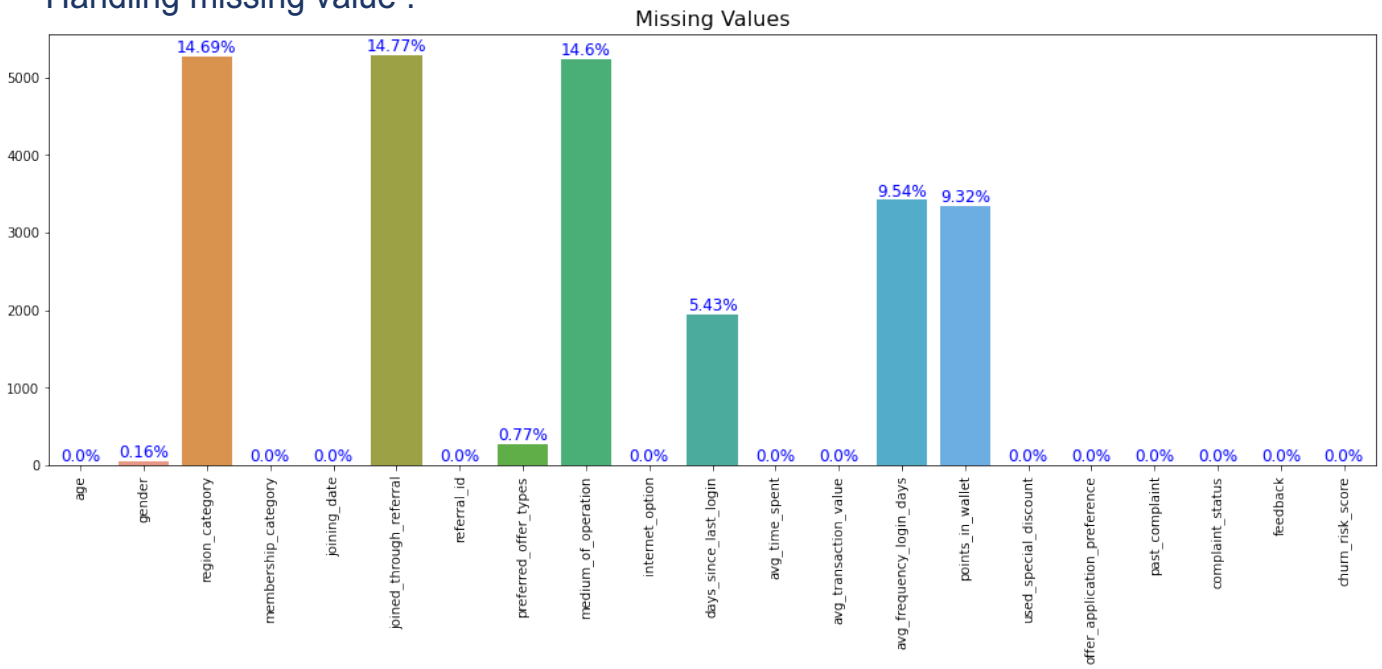
Preprocessing

1. Mengubah nilai ‘?’, ‘Error’, -999 menjadi Mising value (NaN)

2. Menghapus row churn_risk score -1

3. Menyesuaikan Tipe Data Fitur :
   - Tipe data avg_frequency_login_days dari tipe data objek menjadi float
   - Tipe data joining_date dari tipe data objek menjadi datetime

4. Handling missing value :



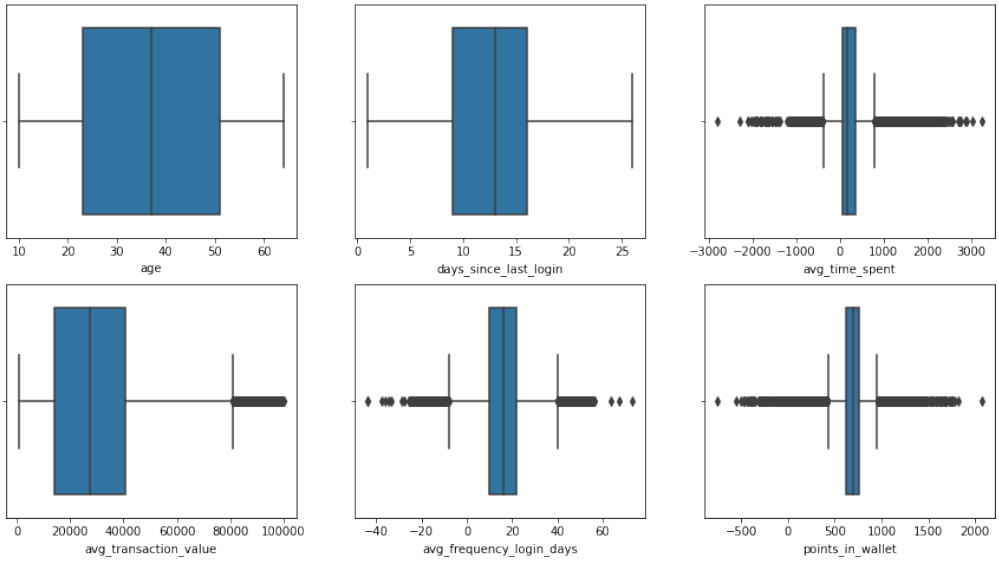
- Missing value di fitur joined_through_referral mengambil informasi dari fitur referral_id

joined_through_referral	referral_id
No	xxxxxxxx

joined_through_referral	referral_id
Yes	CID12313

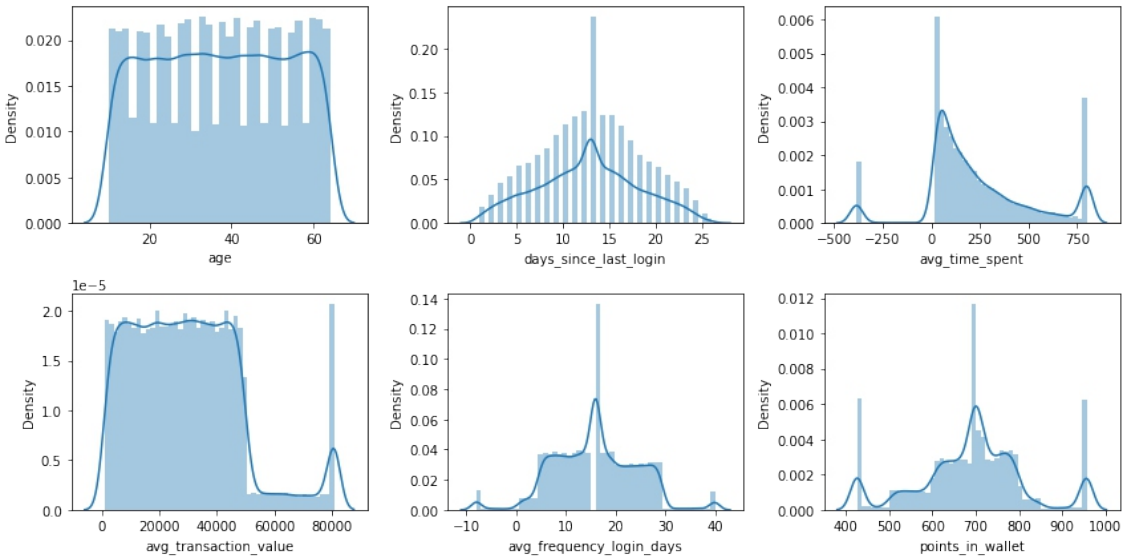
- Missing value di fitur tipe data numerik diganti dengan nilai median dari masing – masing fitur
- Missing value di fitur tipe data kategorik diganti dengan nilai proporsi terbanyak dari masing – masing fitur

5 Handling outlier :



- Untuk masing masing fitur nilai outlier dibawah batas bawah diganti dengan nilai batas bawah
- Untuk masing masing fitur nilai outlier diatas batas atas diganti dengan nilai batas atas

6 Mengubah nilai kurang dari 0 pada fitur 'avg_frequency_login_days' dan 'avg_time_spent' menjadi 0



7 Membuat fitur joined_days dari selisih antara tanggal 16 Maret 2021 dikurangin nilai di fitur days_since_last_login dengan tanggal di fitur joining_date

8 Drop fitur customer_id, Name, security_no, last_visit_time, referral_id, joining_date, days_since_last_login, dan past_complaint

9

- Merubah nilai fitur age jika lebih dari 21 menjadi Adult selain itu menjadi Teen
- Merubah nilai fitur avg_transaction_value jika lebih dari 50000 menjadi High selain itu menjadi Low

10 One hot encoder untuk fitur :

- gender
- region_category
- internet_option
- medium_of_operation
- preferred_offer_types
- joined_through_referral
- used_special_discount
- offer_application_preference
- past_complaint

11 Merubah niliai fitur medium_of_operation_Desktop dan medium_of_operation_Smartphone dari 0 menjadi 1 apabila medium_of_operation_Both 1

medium_of_operation_Both	medium_of_operation_Desktop	medium_of_operation_Smartphone
1	0	0

medium_of_operation_Desktop	medium_of_operation_Smartphone
1	1

12 Drop fitur medium_of_operation_Both dan preferred_offer_types_Without_Offers

preferred_offer_types_Credit/Debit Card Offers	preferred_offer_types_Gift Vouchers/Coupons	preferred_offer_types_Without Offers
0	0	1

preferred_offer_types_Credit/Debit Card Offers	preferred_offer_types_Gift Vouchers/Coupons
0	0

13 Label encoder untuk fitur membership_category dengan nilai

- No Membership:0
- Basic Membership:1
- Silver Membership:2
- Gold Membership:3
- Premium Membership:4
- Platinum Membership:5

14 Label encoder untuk fitur feedback dengan nilai

- Products always in Stock : 1
- Quality Customer Care : 1
- Use Friendly Website : 1
- Reasonable Price : 1
- No Reason Specified : 0
- Poor Website:-1
- Poor Product Quality:-1
- Poor Customer Service:-1
- Too many ads:-1

15 RobustScaler

16 Spliting data train size 0.8, data test size 0.2 , random_state ke 0

17 Balancing data Train dengan SMOTE

Model Development

K-Nearest Neighbor

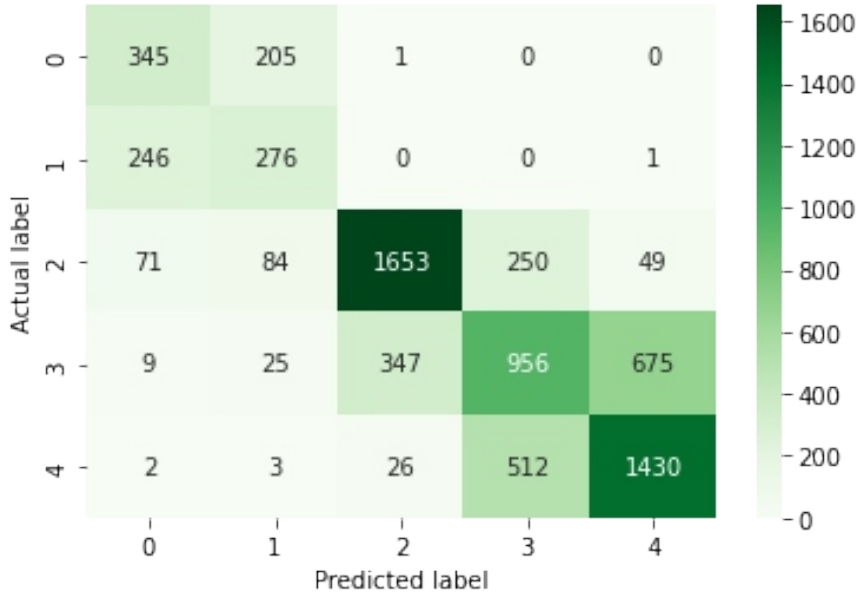
Model KNN untuk data test mendapatkan akurasi sebesar 65,03 % dengan metrik evaluasi :

Report Metrics KNN

	precision	recall	f1-score	support
1	0.51	0.63	0.56	551
2	0.47	0.53	0.49	523
3	0.82	0.78	0.80	2107
4	0.56	0.48	0.51	2012
5	0.66	0.72	0.69	1973
accuracy			0.65	7166
macro avg	0.60	0.63	0.61	7166
weighted avg	0.65	0.65	0.65	7166

F1-score model KNN sebesar 0,65

Confusiun matriks model KNN:



Karena churn risk score 1 dan 2 karesteristiknya berbeda maka untuk data actual churn risk score 1 tetapi diprediksi oleh model KNN menjadi 2 dan sebaliknya tidak masalah. Begitu juga untuk churn risk score 3, 4 dan 5

Dari confussion matrik KNN masih ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model KNN menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tidak tepat karena karesteristik churn risk score 1 dan 2 berbeda dengan churn risk score 3, 4 dan 5.

Decission Tree

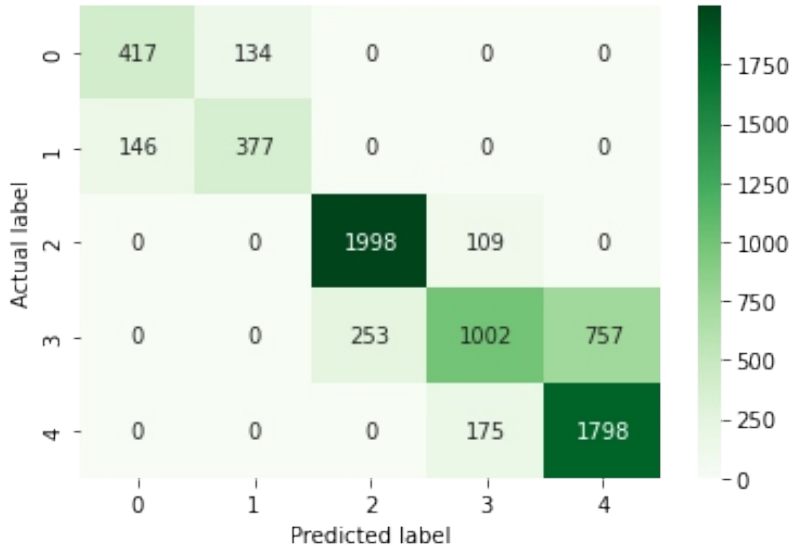
Model Decission Tree untuk data test mendapatkan akurasi sebesar 78,04 % dengan metrik evaluasi :

Report Metrics Decesion Tree

	precision	recall	f1-score	support
1	0.74	0.76	0.75	551
2	0.74	0.72	0.73	523
3	0.89	0.95	0.92	2107
4	0.78	0.50	0.61	2012
5	0.70	0.91	0.79	1973
accuracy			0.78	7166
macro avg	0.77	0.77	0.76	7166
weighted avg	0.78	0.78	0.77	7166

F1-score model Decission Tree sebesar 0,78

Confusiun matriks model Decission Tree:



Karena churn risk score 1 dan 2 karesteristiknya berbeda maka untuk data actual churn risk score 1 tetapi diprediksi oleh model Decission Tree menjadi 2 dan sebaliknya tidak masalah. Begitu juga untuk churn risk score 3, 4 dan 5

Dari confussion matrik Decission Tree sudah tidak ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model Decission Tree menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tepat karena karesteristik churn risk score 1 dan 2 berbeda dengan churn risk score 3, 4 dan 5.

Random Forest

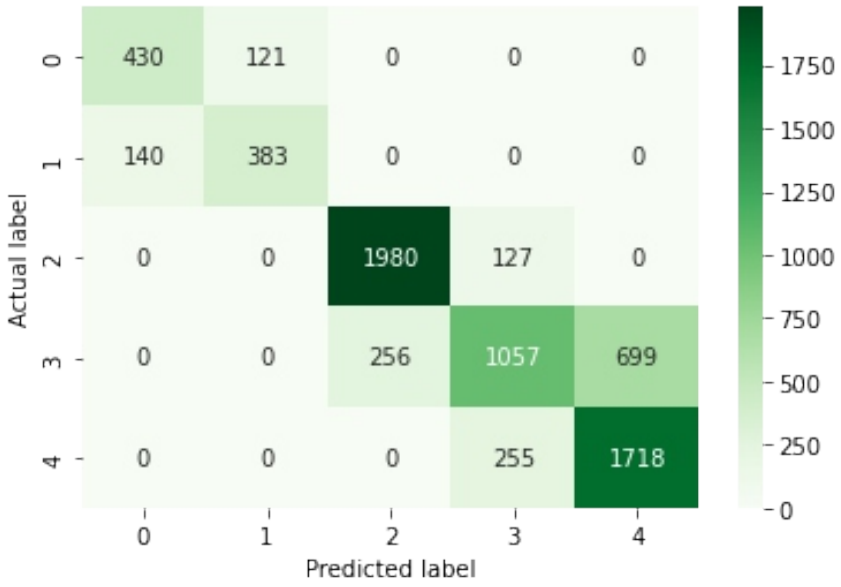
Model Random Forest untuk data test mendapatkan akurasi sebesar 77,70 % dengan metrik evaluasi :

Report Metrics Random Forest

	precision	recall	f1-score	support
1	0.75	0.78	0.77	551
2	0.76	0.73	0.75	523
3	0.89	0.94	0.91	2107
4	0.73	0.53	0.61	2012
5	0.71	0.87	0.78	1973
accuracy			0.78	7166
macro avg	0.77	0.77	0.76	7166
weighted avg	0.78	0.78	0.77	7166

F1-score model Random Forest sebesar 0,78

Confusiun matriks model Random Forest :



Karena churn risk score 1 dan 2 karesteristiknya berbeda maka untuk data actual churn risk score 1 tetapi diprediksi oleh model Random Forest menjadi 2 dan sebaliknya tidak masalah. Begitu juga untuk churn risk score 3, 4 dan 5

Dari confussion matrik Random Forest sudah tidak ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model Random Forest menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tepat karena karesteristik churn risk score 1 dan 2 berbeda dengan churn risk score 3, 4 dan 5.

Support Vector Machine

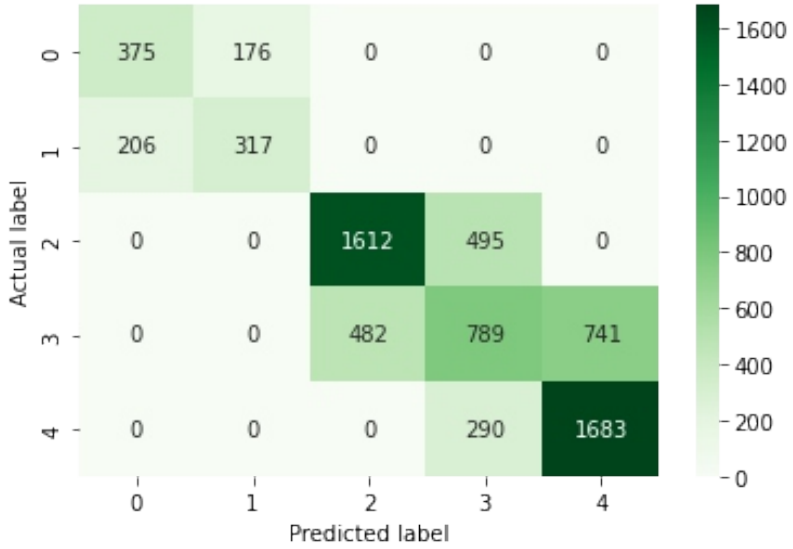
Model SVM untuk data test mendapatkan akurasi sebesar 66,65 % dengan metrik evaluasi :

Report Metrics SVM

	precision	recall	f1-score	support
1	0.65	0.68	0.66	551
2	0.64	0.61	0.62	523
3	0.77	0.77	0.77	2107
4	0.50	0.39	0.44	2012
5	0.69	0.85	0.77	1973
accuracy			0.67	7166
macro avg	0.65	0.66	0.65	7166
weighted avg	0.65	0.67	0.66	7166

F1-score model SVM sebesar 0,67

Confusiun matriks model SVM:



Karena churn risk score 1 dan 2 karesteristiknya berbeda maka untuk data actual churn risk score 1 tetapi diprediksi oleh model SVM menjadi 2 dan sebaliknya tidak masalah. Begitu juga untuk churn risk score 3, 4 dan 5

Dari confussion matrik SVM masih ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model SVM menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tidak tepat karena karesteristik churn risk score 1 dan 2 dengan churn risk score 3, 4 dan 5.

XGBoost

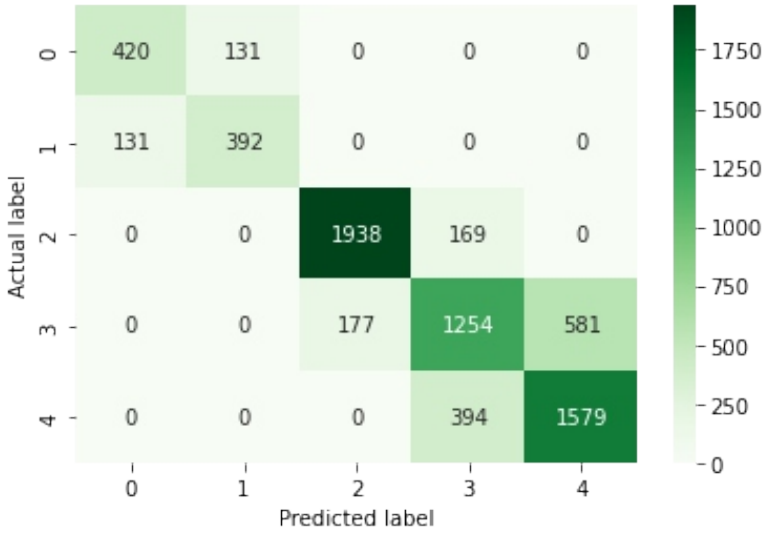
Model XGBoost untuk data test mendapatkan akurasi sebesar 77,91 % dengan metrik evaluasi :

Report Metrics xgb

	precision	recall	f1-score	support
1	0.76	0.76	0.76	551
2	0.75	0.75	0.75	523
3	0.92	0.92	0.92	2107
4	0.69	0.62	0.66	2012
5	0.73	0.80	0.76	1973
accuracy			0.78	7166
macro avg	0.77	0.77	0.77	7166
weighted avg	0.78	0.78	0.78	7166

F1-score model XGBoost sebesar 0,78

Confusiun matriks model XGBoost :



Karena churn risk score 1 dan 2 karesteristiknya berbeda maka untuk data actual churn risk score 1 tetapi diprediksi oleh model XGBoost menjadi 2 dan sebaliknya tidak masalah. Begitu juga untuk churn risk score 3, 4 dan 5

Dari confussion matrik XGBoost sudah tidak ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model XGBoost menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tepat karena karesteristik churn risk score 1 dan 2 berbeda dengan churn risk score 3, 4 dan 5.

Conclussion

Insight untuk Bisnis

- Bisnis perlu memerhatikan complaint customer jangan sampai ada yang di abaikan dan bursaha memperbaiki berdasarkan complaint tersebut. Churn rate akan tinggi ketika ada complaint dari customer diabaikan atau tidak diperhatikan oleh bisnins
- Bisnis perlu fokus memperbaiki pelayanan (service) terhadap customer dengan churn risk score 3, 4 dan 5 karena ada yang memberikan feedback negatif.
- Bisnis peelu mengevaluasi mengapa customer yang sudah menjadi membership (Basic Membership - Platinum Membership) masih memiliki feedback negatif atau complaint. Bisnis harus fokus memperbaiki pelayan terhadap mereka.

Rekomendasi Model

- Model yang akan direkomendasikan untuk memprediksi churn risk score adalah Model Decission Tree karena selain memiliki akurasi tertinggi yaitu 78,04% juga berdasarkan Evaluation matriks memiliki f1-score 0,78 dan confussion matriks yang tepat tidak ada untuk data aktual churn risk score 1 atau 2 di prediksi oleh model Decission Tree menjadi churn risk score 3 atau 4 atau 5. Begitu juga sebaliknya. Ini tepat karena karesteristik churn risk score 1 dan 2 berbeda dengan churn risk score 3, 4 dan 5.