

ANALISA PREDIKSI NASABAH BANK BERHENTI



Siti Farichatun Dharmaningsih
DS Batch 23B

Background

Menemukan hal-hal yang menyebabkan nasabah Bank keluar dan mencoba membuat model prediksi atas nasabah-nasabah yang akan keluar.

Goal

Tujuan analisa ini adalah untuk mencari faktor yang mempengaruhi nasabah bank yang keluar (Exited). Kemudian dengan mempertimbangkan semua faktor, akan dipersiapkan model yang akurat, yang akan memprediksi nasabah yang keluar.



DATA

Data yang digunakan:

Bank Customer Churn

(source: <https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/code>)

Dataset ini berisi catatan nasabah bank yang keluar, meliputi kolom-kolom: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited, Complain, Satisfaction Score, Card Type, Point Earned.

Data Exploration and Cleaning:

Mendapatkan informasi dasar terhadap dataset seperti jumlah baris dan kolom, tipe data tiap kolom dan sebagainya. Selanjutnya juga melakukan pembersihan data dari duplikat maupun missing value, agar dataset siap digunakan.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   RowNumber              10000 non-null  int64
1   CustomerId             10000 non-null  int64
2   Surname                10000 non-null  object
3   CreditScore             10000 non-null  int64
4   Geography              10000 non-null  object
5   Gender                 10000 non-null  object
6   Age                   10000 non-null  int64
7   Tenure                 10000 non-null  int64
8   Balance                10000 non-null  float64
9   NumOfProducts          10000 non-null  int64
10  HasCrCard              10000 non-null  int64
11  IsActiveMember         10000 non-null  int64
12  EstimatedSalary         10000 non-null  float64
13  Exited                 10000 non-null  int64
14  Complain               10000 non-null  int64
15  Satisfaction Score     10000 non-null  int64
16  Card Type              10000 non-null  object
17  Point Earned           10000 non-null  int64
dtypes: float64(2), int64(12), object(4)
memory usage: 1.4+ MB
```

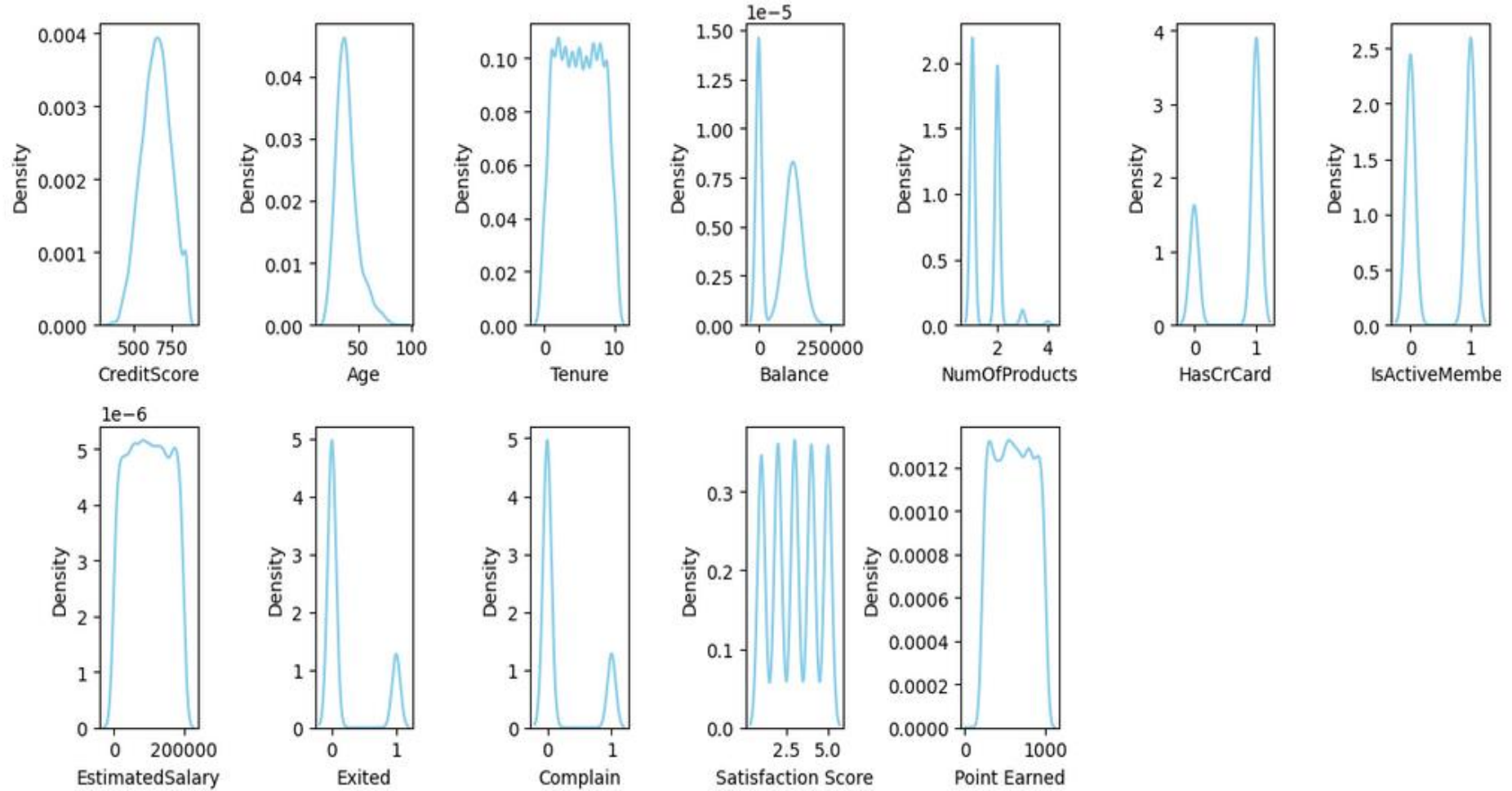
RowNumber	10000
CustomerId	10000
Surname	2932
CreditScore	460
Geography	3
Gender	2
Age	70
Tenure	11
Balance	6382
NumOfProducts	4
HasCrCard	2
IsActiveMember	2
EstimatedSalary	9999
Exited	2
Complain	2
Satisfaction Score	5
Card Type	4
Point Earned	785
dtype: int64	

Terdari dari:

- 10.000 baris
- 18 kolom

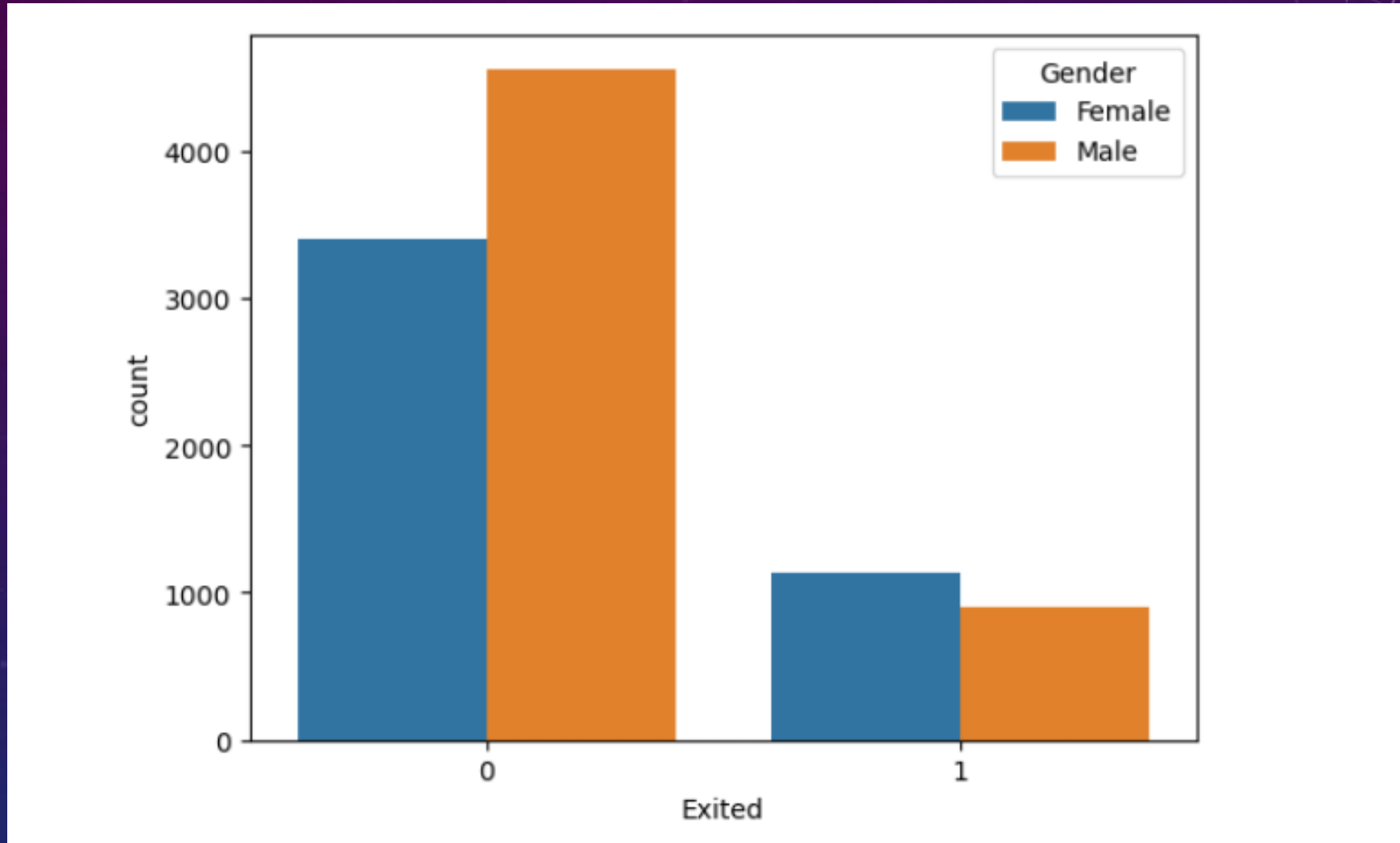
EXPLORATORY DATA ANALYSIS





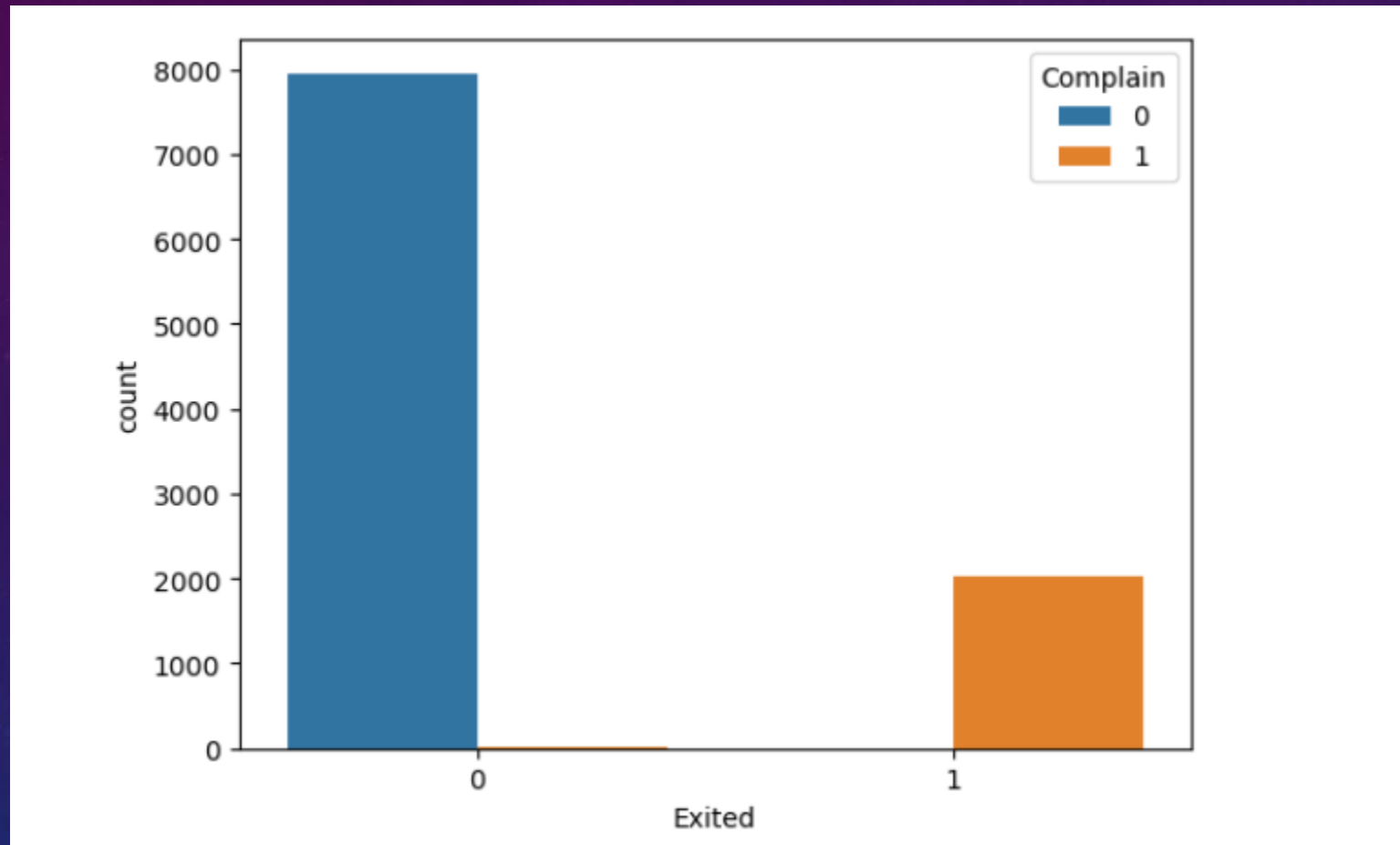
KDE Plot: memperlihatkan distribusi dari masing-masing feature

Bivariate Analysis



- * Secara umum, jumlah nasabah yang keluar adalah 20% dari total nasabah.
- * Perbedaan jumlah Female jauh lebih tinggi dari Male untuk nasabah yang '0' (tidak keluar).
- * Jumlah Male sedikit lebih banyak dari Female untuk nasabah yang '1' (keluar).

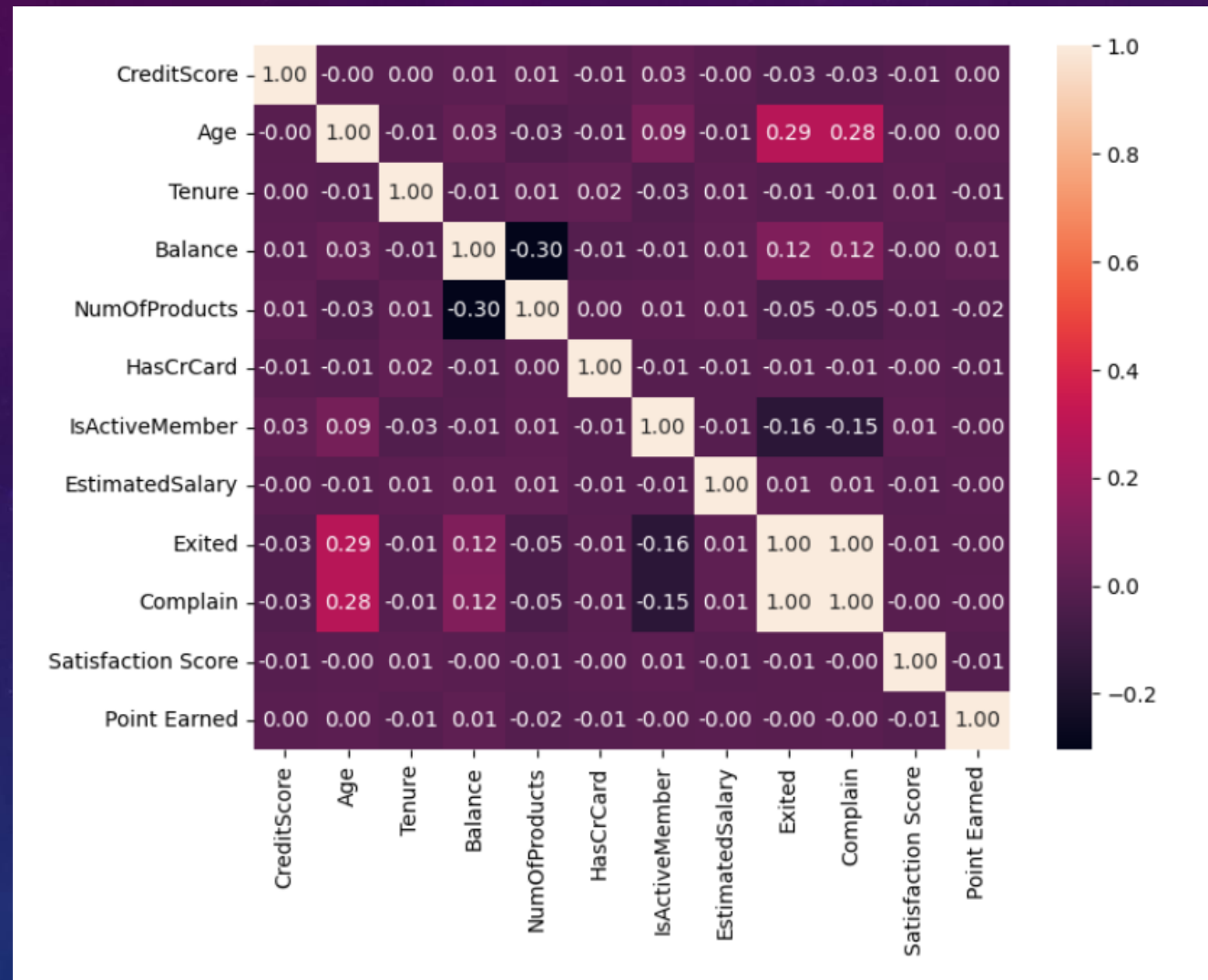
Distribusi nasabah keluar berdasarkan 'Complain'



Sebaran nasabah keluar berdasarkan 'Complain';

* Dari 20% jumlah nasabah yang mengajukan complain, hampir seluruhnya akhirnya keluar.

Heatmap



Memperlihatkan hubungan yang kuat pada feature 'Age', 'Exited' dan 'Complain'.

Agregasi nasabah keluar
berdasarkan Complain

Complain		Exited count
0	4	
1	2034	

Agregasi nasabah keluar berdasarkan
CreditCard dan Gender

HasCrCard		Gender	Exited count
0	0	0	344
		1	269
1	0	0	795
		1	630

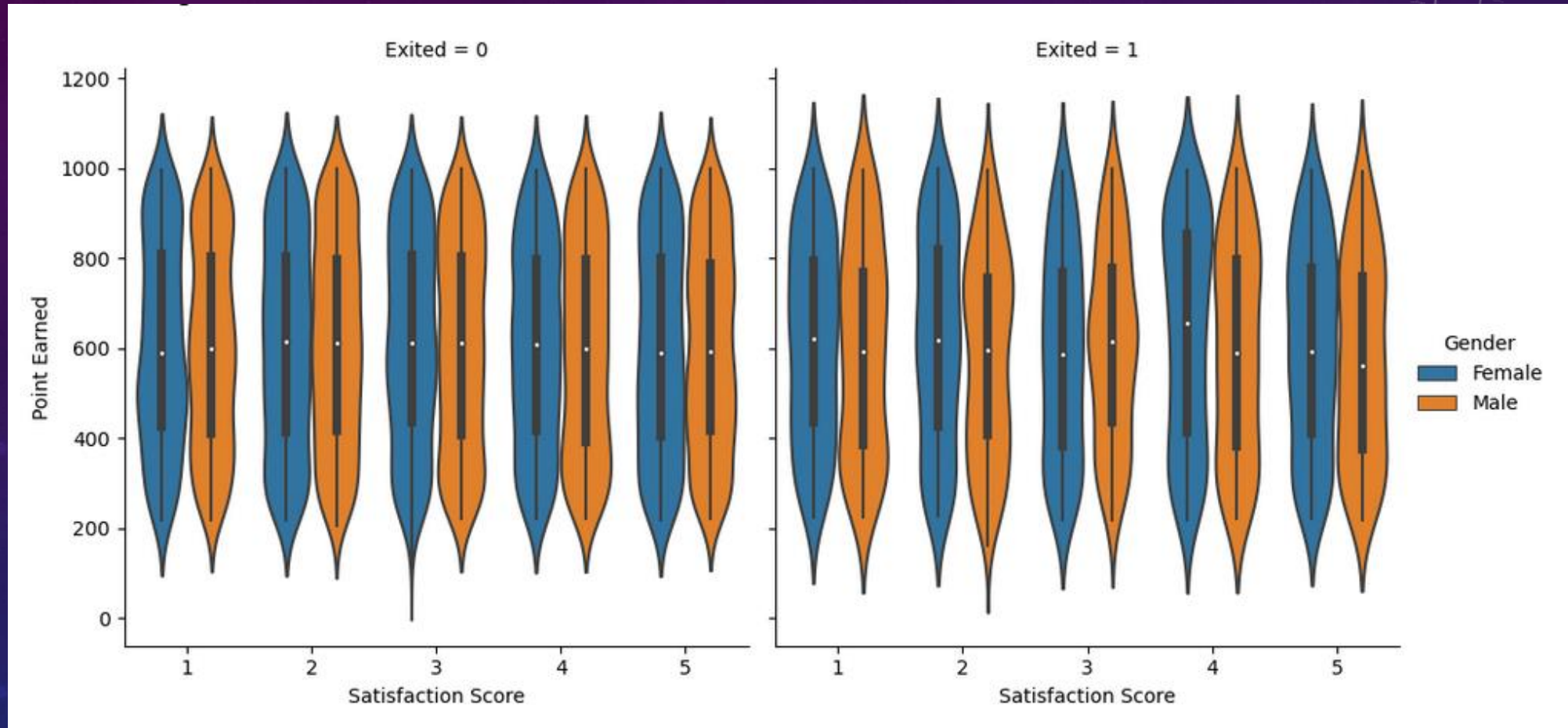
Agregasi nasabah keluar berdasarkan
Geography dan Gender

		Exited count
Geography	Gender	
France	0	460
	1	351
Germany	0	448
	1	366
Spain	0	231
	1	182

Agregasi nasabah Complain berdasarkan
Exited dan Gender

		Complain count
Exited	Gender	
0	Female	5
	Male	5
1	Female	1137
	Male	897

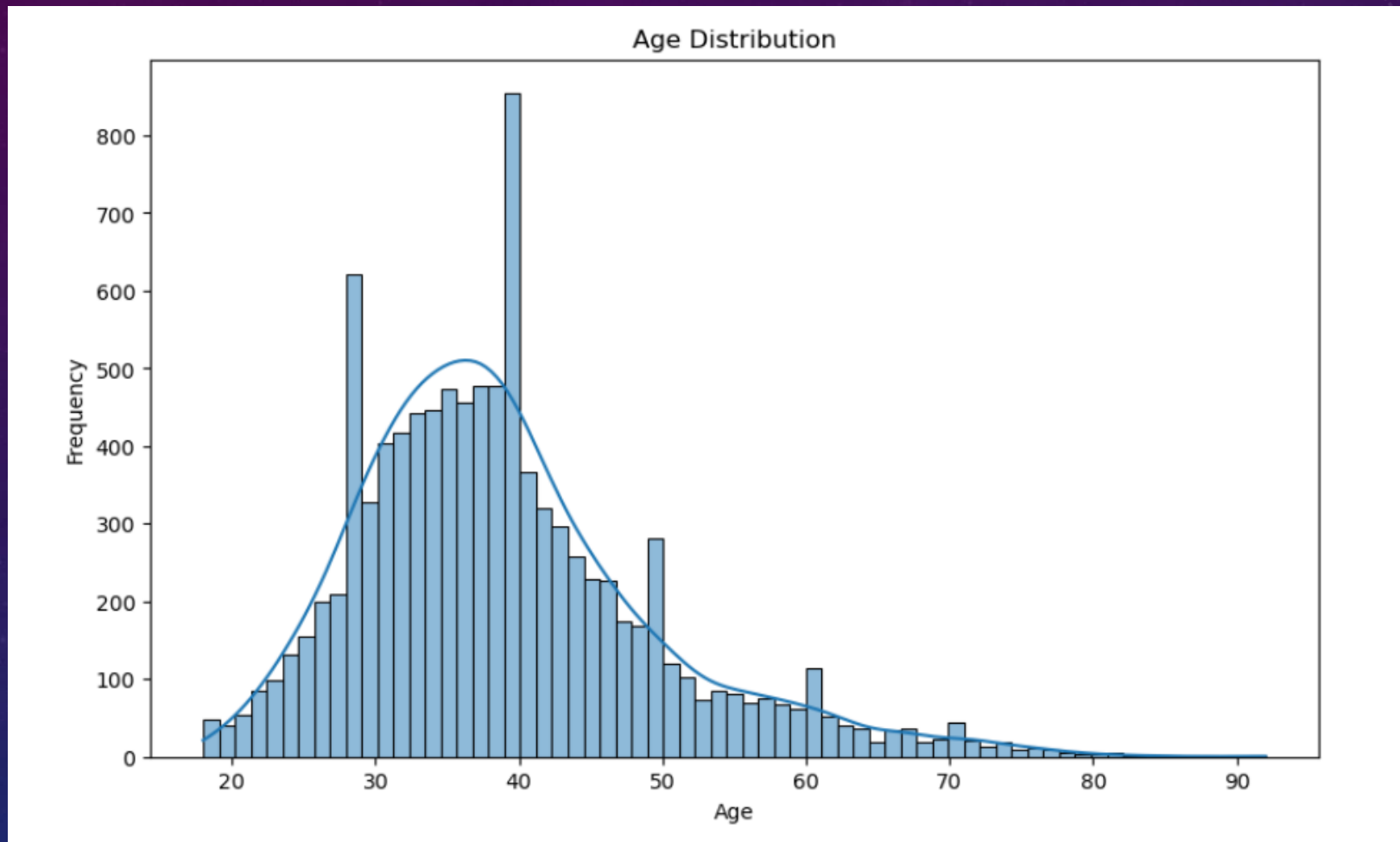
Distribusi nasabah keluar berdasarkan Satisfaction Score



Sebaran distribusi nasabah keluar berdasarkan 'Satisfaction Score':

* Satisfaction Score tidak mempengaruhi keputusan nasabah untuk berhenti.

Sebaran usia nasabah



Sebaran usia nasabah:

- Nasabah terbanyak di rentang usia 29 s/d 45 tahun, dengan jumlah terbanyak pada usia 29 tahun (600 orang) dan 40 tahun (850 orang).

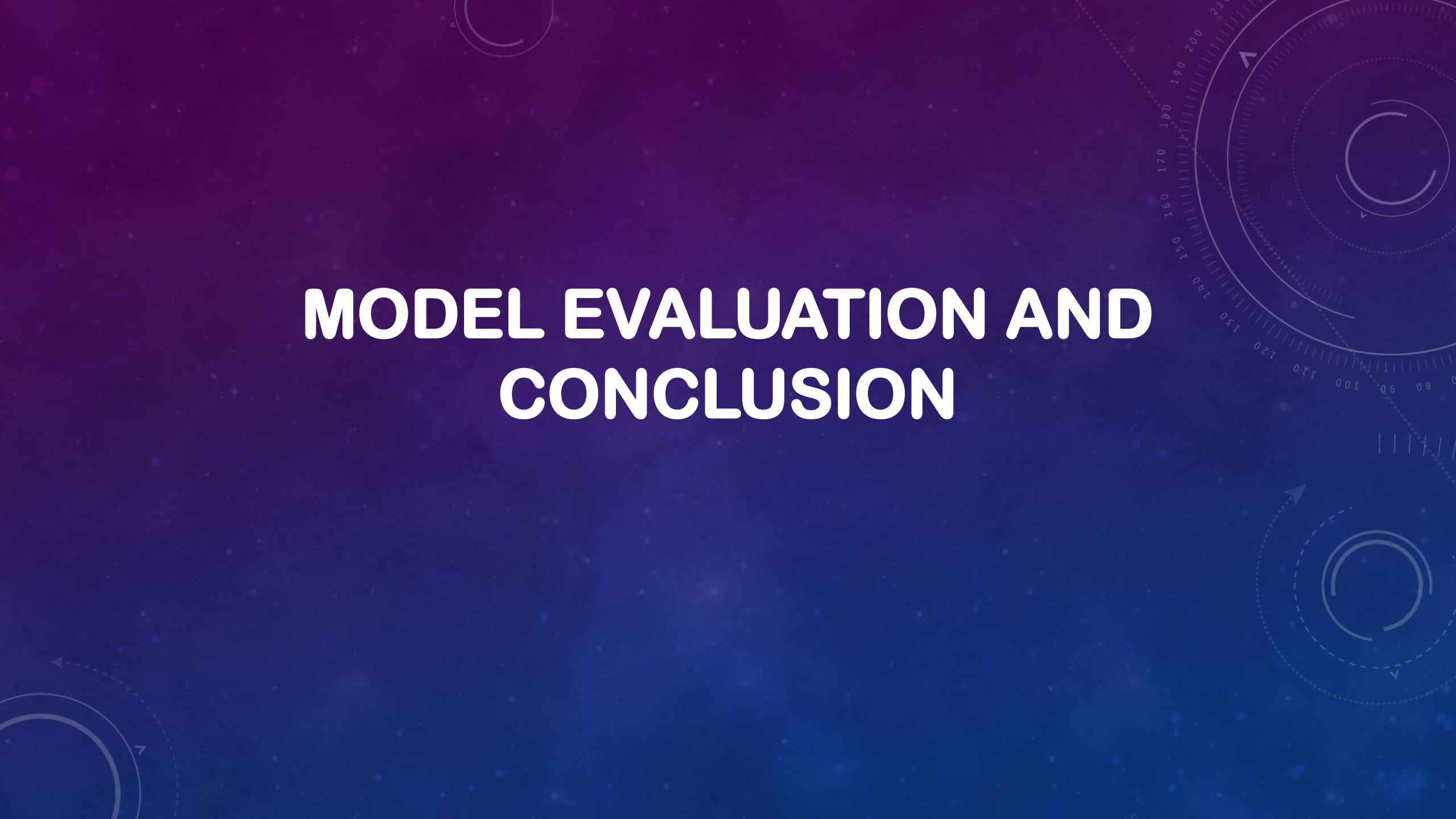
INITIAL CONCLUSION

Initial Conclusion:

Merujuk tujuan awal dilakukan Analisa, maka temuan awal sebagai berikut:

- Dari 10000 orang nasabah, sejumlah 2038 orang keluar (20%).
- Dari 2038 orang yang keluar, sejumlah 2034 orang mengajukan complain.
- Dari 2034 orang nasabah yang keluar dan complain, sejumlah 1142 orang adalah Female.

MODEL EVALUATION AND CONCLUSION



	Algoritma	ROC_AUC_Mean	ROC_AUC_Std	Recall_Mean	Recall_Std	Precision_Mean	Precision_Std	F1_score_Mean	F1_score_Std
1	Random Forest	99.92	0.07	99.75	0.09	99.56	0.24	99.66	0.17
4	Decision Tree Classifier	99.29	0.41	98.71	0.79	99.44	0.27	99.07	0.46
2	SVM	79.28	0.36	23.16	0.75	99.73	0.38	37.58	1.00
3	KNN	69.44	1.16	21.58	2.14	48.89	2.36	29.91	2.40
5	Gaussian NB	80.76	0.11	9.13	0.75	46.92	7.96	15.24	1.35
0	Logistic Regression	30.19	1.17	0.00	0.00	0.00	0.00	0.00	0.00

Menggunakan model dengan Algoritma:

- Logistic Regression
- Decision Tree Classifier
- Gaussian NB
- SVM
- KNN
- Random Forest

Memperlihatkan **hasil terbaik** adalah dengan menggunakan **model Random Forest** dengan nilai **ROC_AUC_Mean sebesar 99.92%**.

Evaluation for Data Test – Random Forest

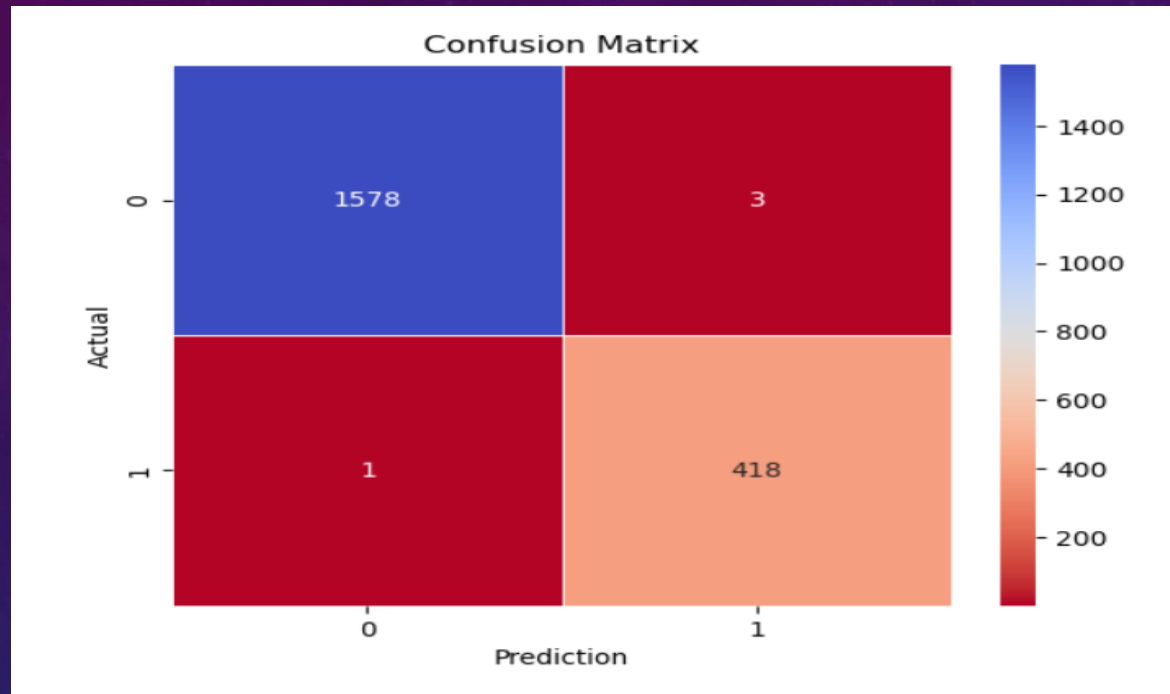
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1581
1	0.99	1.00	1.00	419
accuracy			1.00	2000
macro avg	1.00	1.00	1.00	2000
weighted avg	1.00	1.00	1.00	2000

Intepretasi:

- **Precision:** Untuk **kelas 0** (nasabah tidak keluar), **precision adalah 1.00**, yang berarti 100% dari hasil yang diprediksi sebagai kelas 0 (nasabah tidak keluar) adalah benar-benar kelas 0. Sedangkan untuk **kelas 1** (nasabah keluar), **precision adalah 0.99**, yang berarti hampir semua dari hasil yang diprediksi sebagai kelas 1 (nasabah keluar) adalah benar-benar kelas 1.
- **Recall:** Untuk **kelas 0**, **recall adalah 1.00**, yang berarti seluruh kasus kelas 0 berhasil diidentifikasi dengan benar oleh model. Demikian juga untuk **kelas 1**, **recall adalah 1.00**, yang berarti semua kasus kelas 1 berhasil diidentifikasi dengan benar oleh model.
- **F1-score:** Untuk **kelas 0 dan kelas 1**, kedua-duanya memiliki **F1-score sebesar 1.00**.
- **Support:** terdapat 1581 sampel untuk kelas 0 dan 419 sampel untuk kelas 1.
- **Accuracy:** **Akurasi adalah 1.00**, yang berarti semua prediksi yang dilakukan oleh model adalah benar.

Dari data ini, kita dapat menyimpulkan bahwa model memiliki kinerja yang sangat baik dan mampu membuat prediksi yang sangat baik.

Confusion Matrix for Random Forest

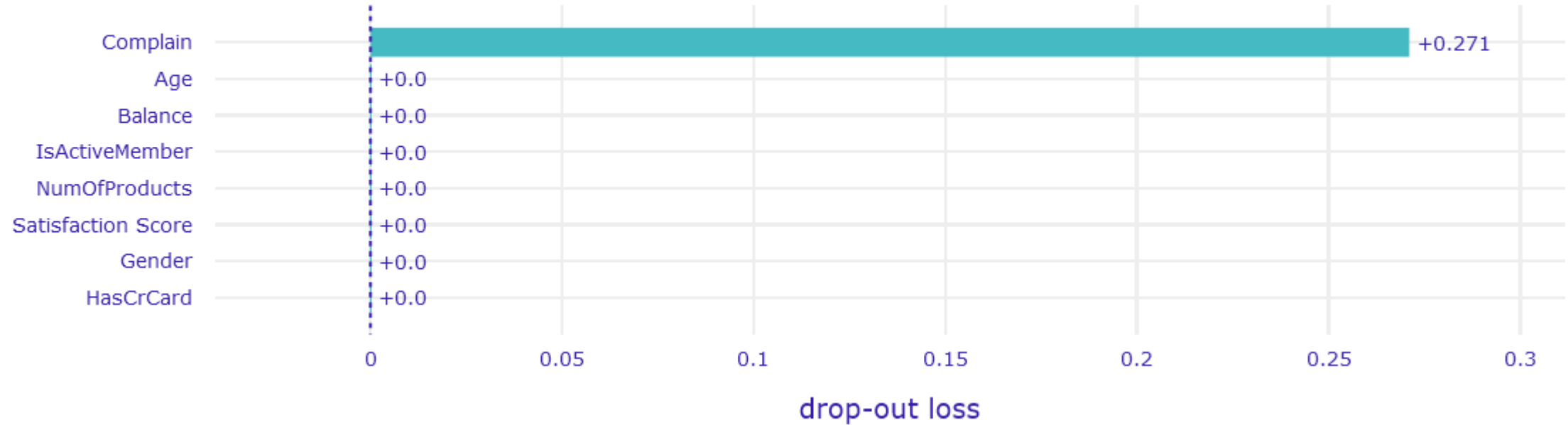


Dari **confusion matrix** diatas, kita memiliki informasi berikut:

- **True Positive (TP)** adalah **418**, yang merupakan jumlah nasabah yang diprediksi memiliki keluar dan memang keluar.
- **True Negative (TN)** adalah **1578**, yang merupakan jumlah nasabah yang diprediksi tidak keluar dan memang tidak keluar.
- **False Positive (FP)** adalah **3**, yang merupakan jumlah nasabah yang diprediksi keluar, tetapi sebenarnya tidak keluar.
- **False Negative (FN)** adalah **1**, yang merupakan jumlah nasabah yang diprediksi tidak keluar, tetapi sebenarnya keluar.

Variable Importance

Random Forest Interpretation



Hasil Analisa memperlihatkan bahwa Complain merupakan factor utama penyebab nasabah keluar, sehingga 100% dipastikan bahwa nasabah yang mengajukan complain akan berhenti.



INSIGHT


Insight

1. Pembuatan model memperkuat kesimpulan awal hasil dari EDA yaitu dari 2038 nasabah yang keluar, hampir seluruhnya mengajukan complain, sebanyak 2034.
2. Model memiliki keberhasilan amat tinggi/baik yaitu sebesar 99.85% dalam memprediksi customer keluar.
3. Faktor utama yang mengakibatkan kemungkinan nasabah keluar adalah 'Complain'.
4. Pada EDA diketahui bahwa hampir 60% nasabah keluar adalah Female, sebanyak 1142 orang.
5. Feature lain terlihat adalah 'Age', dimana hal ini dimungkinkan karena nasabah yang berumur bisa saja meninggal atau memutuskan berhenti karena mobilitas yang semakin sulit untuk ke Bank.
6. Feature lain yang juga terlihat adalah 'Balance', dimana alasan yang masuk akal bahwa nasabah yang memiliki saldo yang sedikit akhirnya memutuskan berhenti.

RECOMMENDATION

Recommendation

1. Memperkuat Sistem Complain – meningkatkan responsivitas dan efektifitas dalam menangani keluhan nasabah serta memberikan solusi yang memuaskan. Melakukan evaluasi atas seluruh complain yang diajukan untuk dibuatkan kebijakan baru dalam rangka perbaikan.
2. Penyesuaian Layanan – mengadakan layanan dan strategi baru melalui segmentasi nasabah berdasarkan karakteristik demografis termasuk usia dan jenis kelamin.
3. Analisis Pelanggan yang Rentan Berhenti – melakukan Analisa lebih dalam mengenai feature 'Age' dan 'Balance', mempertimbangkan kemungkinan untuk membuat layanan baru yang bisa mengurangi nasabah berhenti dari feature ini.
4. Strategi Layanan Terpersonalisasi – mengembangkan strategi untuk mempertahankan nasabah terpersonalisasi, dengan penawaran khusus, loyalty reward maupun layanan tambahan lain yang disesuaikan dengan preferensi individu nasabah. Tujuannya adalah menambah keterikatan nasabah terhadap Bank.
5. Monitoring dan Evaluasi Berkelanjutan – melakukan monitoring dan evaluasi secara berkala terhadap efektivitas strategi retensi yang diimplementasikan. Dengan melakukan evaluasi berkelanjutan, Bank dapat mengidentifikasi strategi yang efektif dan melakukan penyesuaian yang diperlukan sesuai dengan perubahan dalam perilaku nasabah atau kondisi pasar.



Semoga memberikan manfaat dalam memperluas wawasan.

Terima Kasih.