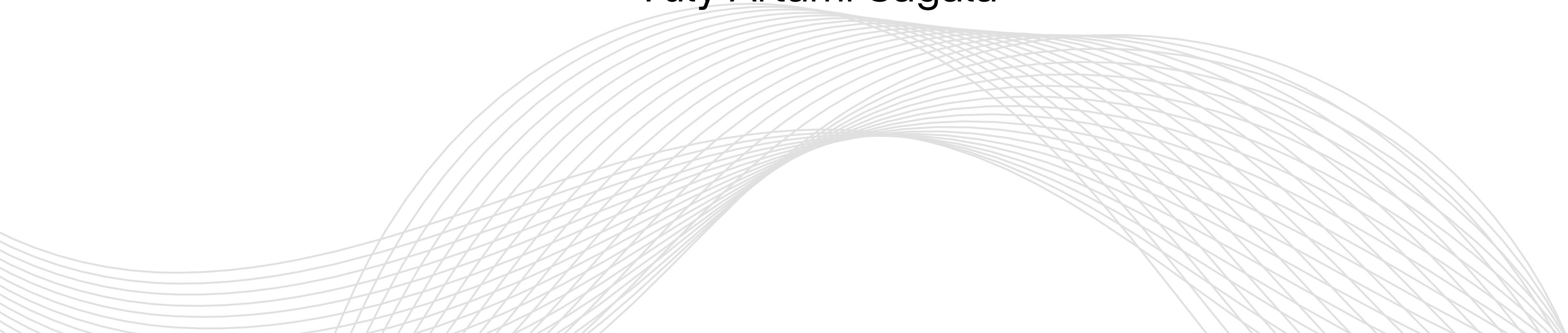




HOME CREDIT SCORECARD

Project-Based Internship: Data Scientist
Home Credit Indonesia x Rakamin Academy
Yuly Artami Sagala



Background

- Home Credit Indonesia merupakan perusahaan pembiayaan berbasis teknologi.
- Home Credit Indonesia ingin memastikan pelanggan yang mampu melakukan pelunasan tidak ditolak ketika melakukan pengajuan pinjaman

Objektive

- Analisis perilaku nasabah
- Memprediksi nasabah berisiko dan tidak berisiko
- Membangun model prediktif yang dapat mengidentifikasi nasabah yang berisiko gagal bayar

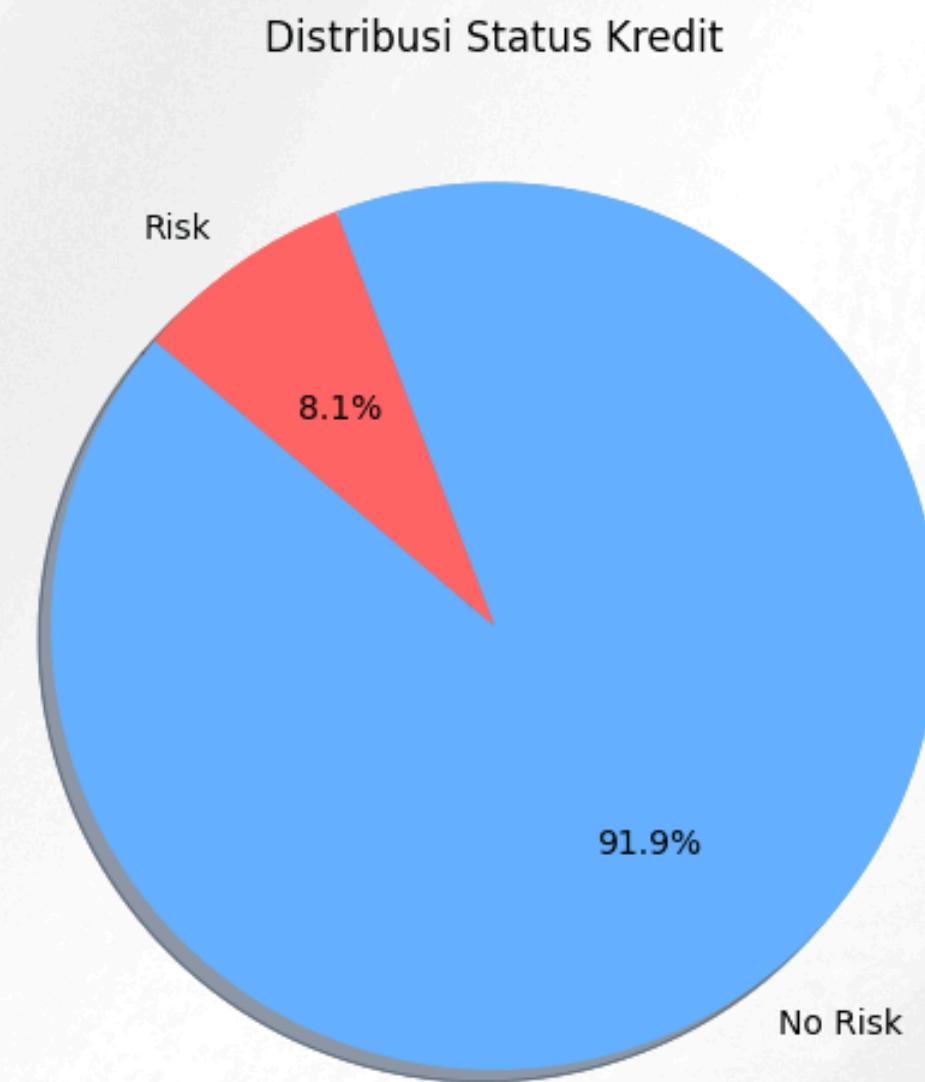
Table of contents

1	Data Understanding	5	Feature Selection
2	Data Preprocessing	6	Modelling and Evaluation
3	Exploratory Data Analysis	7	Prediction Result
4	Feature Engineering	8	Conclusion and Recomendation



Data Understanding

- Dataset yang digunakan pada proyek ini adalah “application_train.csv”
- Kolom “TARGET” menunjukkan label dari masalah, dengan 0 = nasabah tidak berisiko dan 1 = nasabah berisiko gagal bayar
- Dataset terdiri dari 307511 baris dengan nilai unik berdasarkan ID pinjaman dalam “SK_ID_CURR”
- Dataset terdiri dari 122 kolom fitur yang terkait dengan demografi nasabah
- Terdapat 106 data numerik dan 16 data kategorik



Data Preprocessing

```
jumlah_missing = df.isnull().any().sum()  
print(jumlah_missing)  
  
67
```

Terdapat 67 kolom yang terdapat missing values. Missing values diatasi dengan menghapus kolom yang memiliki missing value > 50% dan menggunakan modus dan median.

```
df.duplicated().sum()  
  
np.int64(0)
```

Tidak ada duplikat pada dataset

```
df = df.drop(columns = ['SK_ID_CURR'])
```

Menghapus kolom yang tidak diperlukan

```
FLAG_DOCUMENT = [col for col in df.columns if 'FLAG_DOCUMENT' in col]  
df.drop(columns = FLAG_DOCUMENT, axis=1, inplace=True)
```

Data Preprocessing

Kolom yang mengandung 'XNA' atau 'Unknown':
['CODE_GENDER', 'NAME_FAMILY_STATUS', 'ORGANIZATION_TYPE']

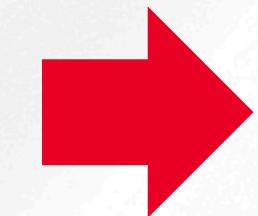
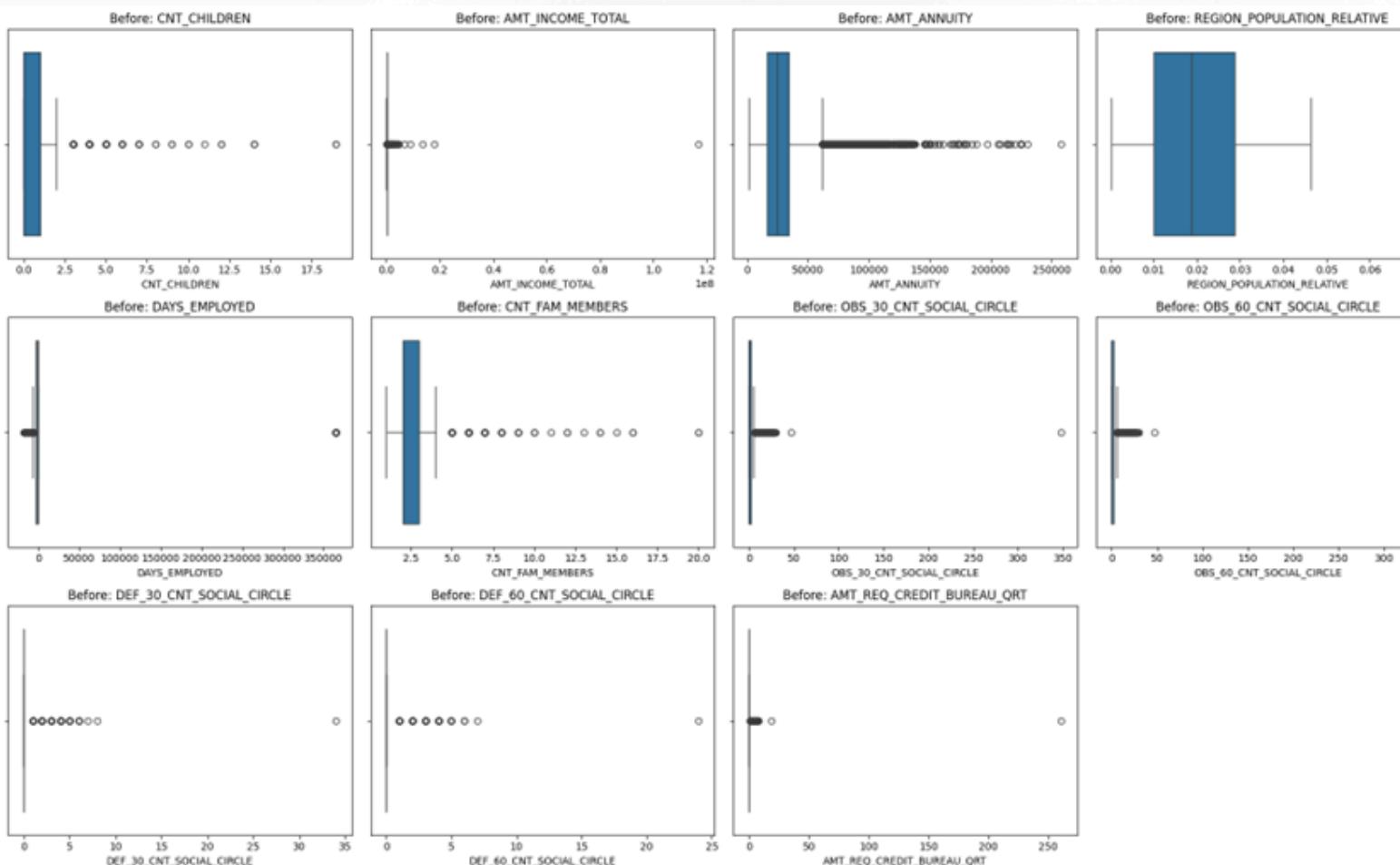
Terdapat 3 kolom yang mengandung 'XNA' atau 'Unknown' pada data yang ditangani menggunakan modus

Kolom dengan lebih dari 10 nilai unik (high cardinality):
['OCCUPATION_TYPE', 'ORGANIZATION_TYPE']

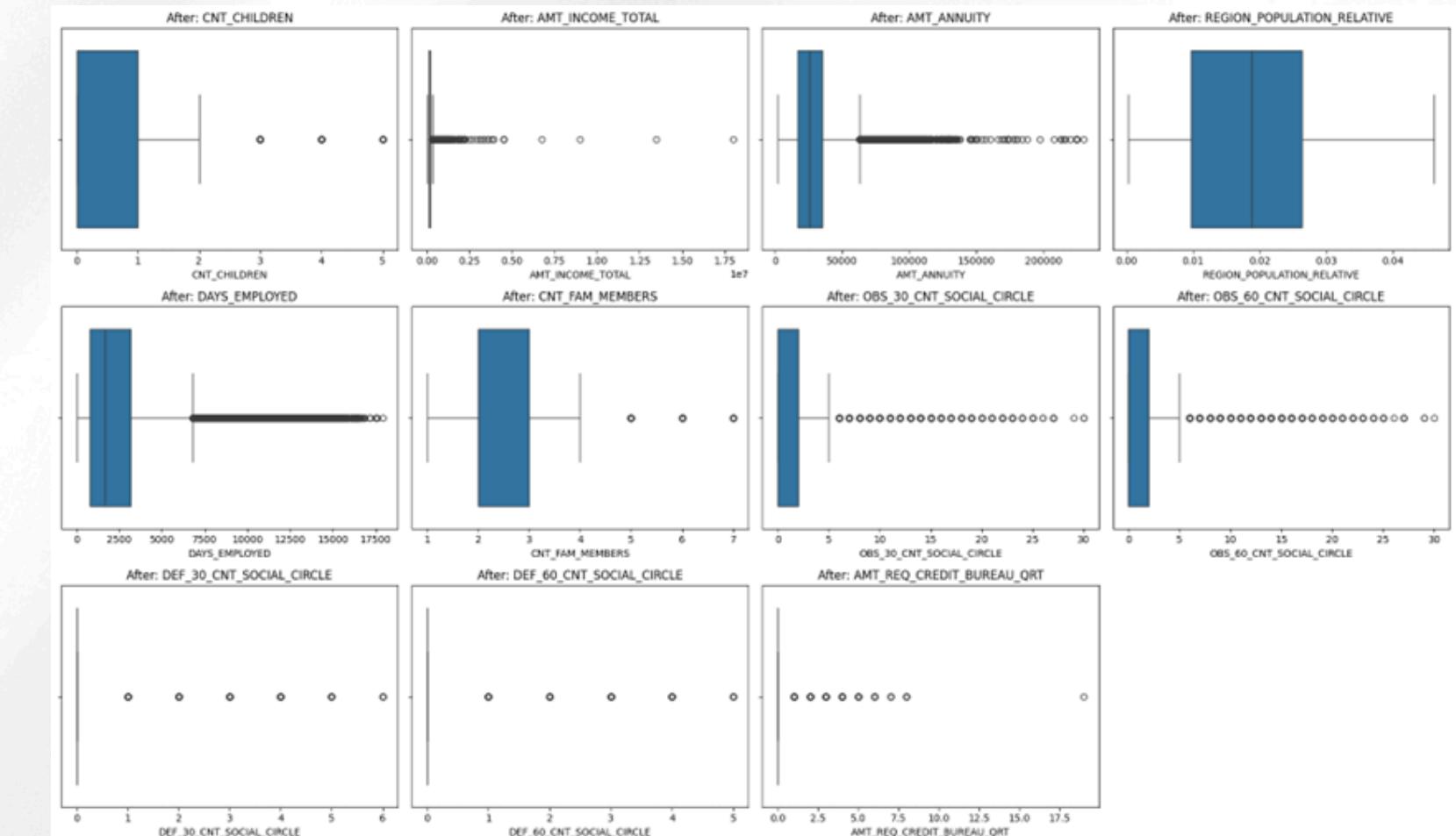
Menghapus kolom yang memiliki lebih dari 10 nilai unik.

Handling Outliers

Before

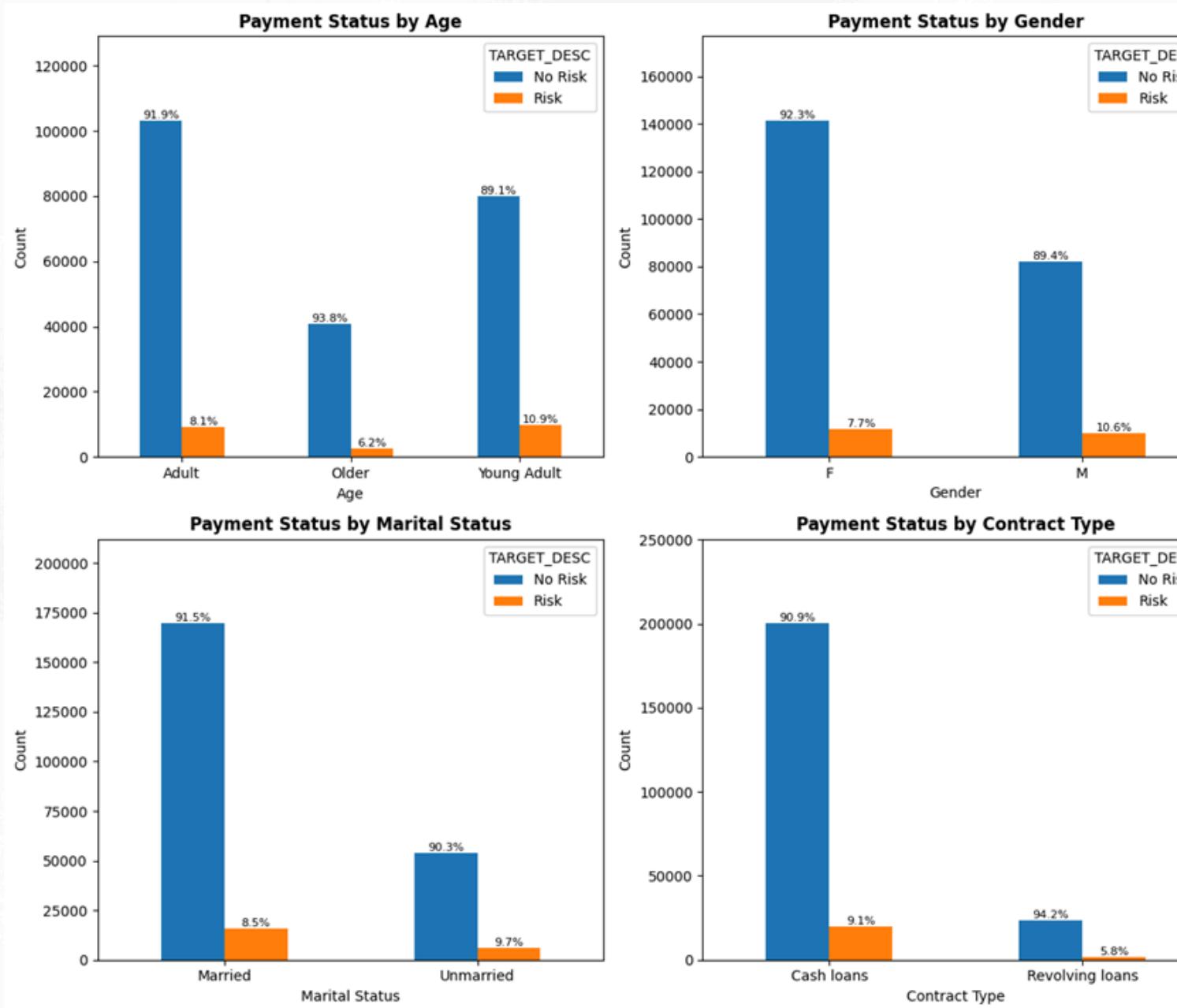


After



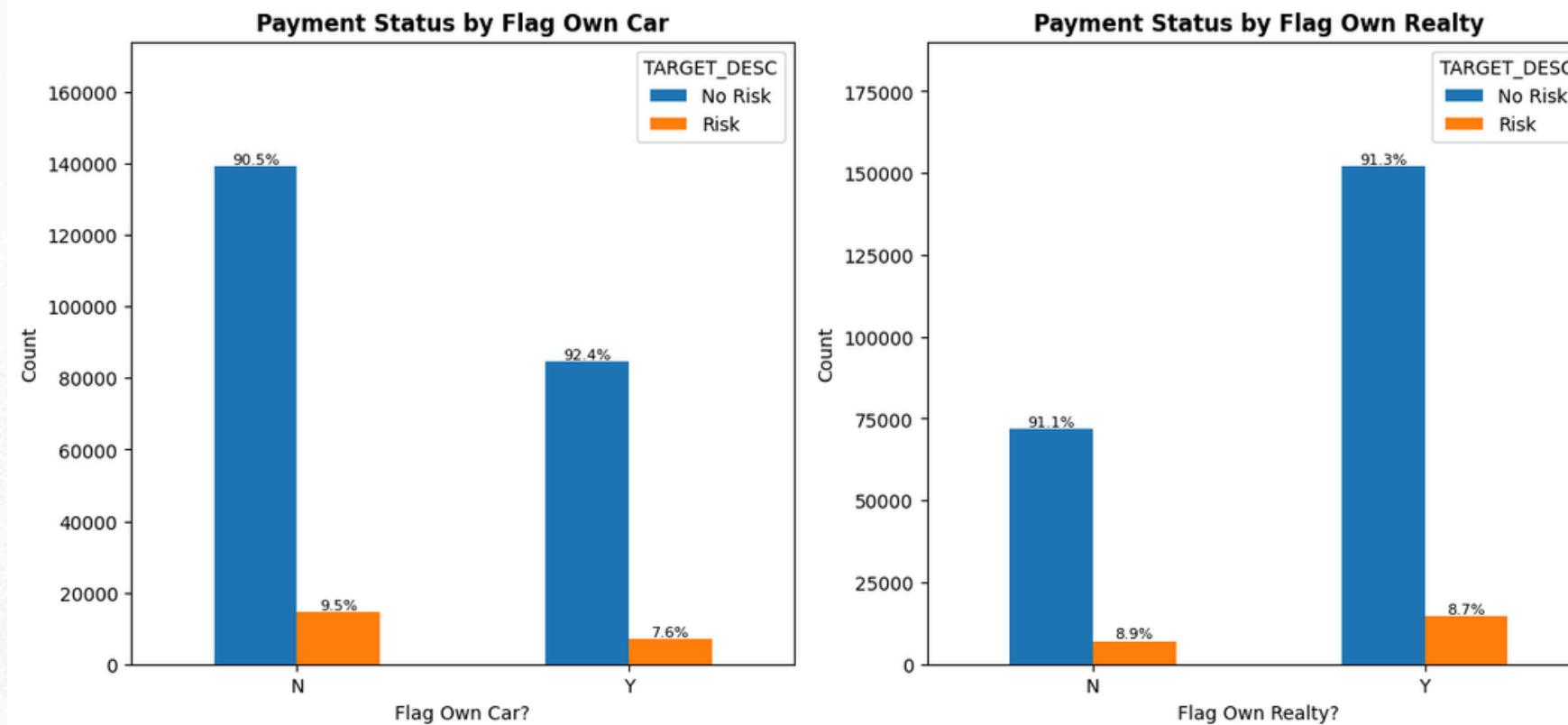
Menghapus baris data yang mengandung outlier dengan nilai ekstrem.

Exploratory Data Analysis



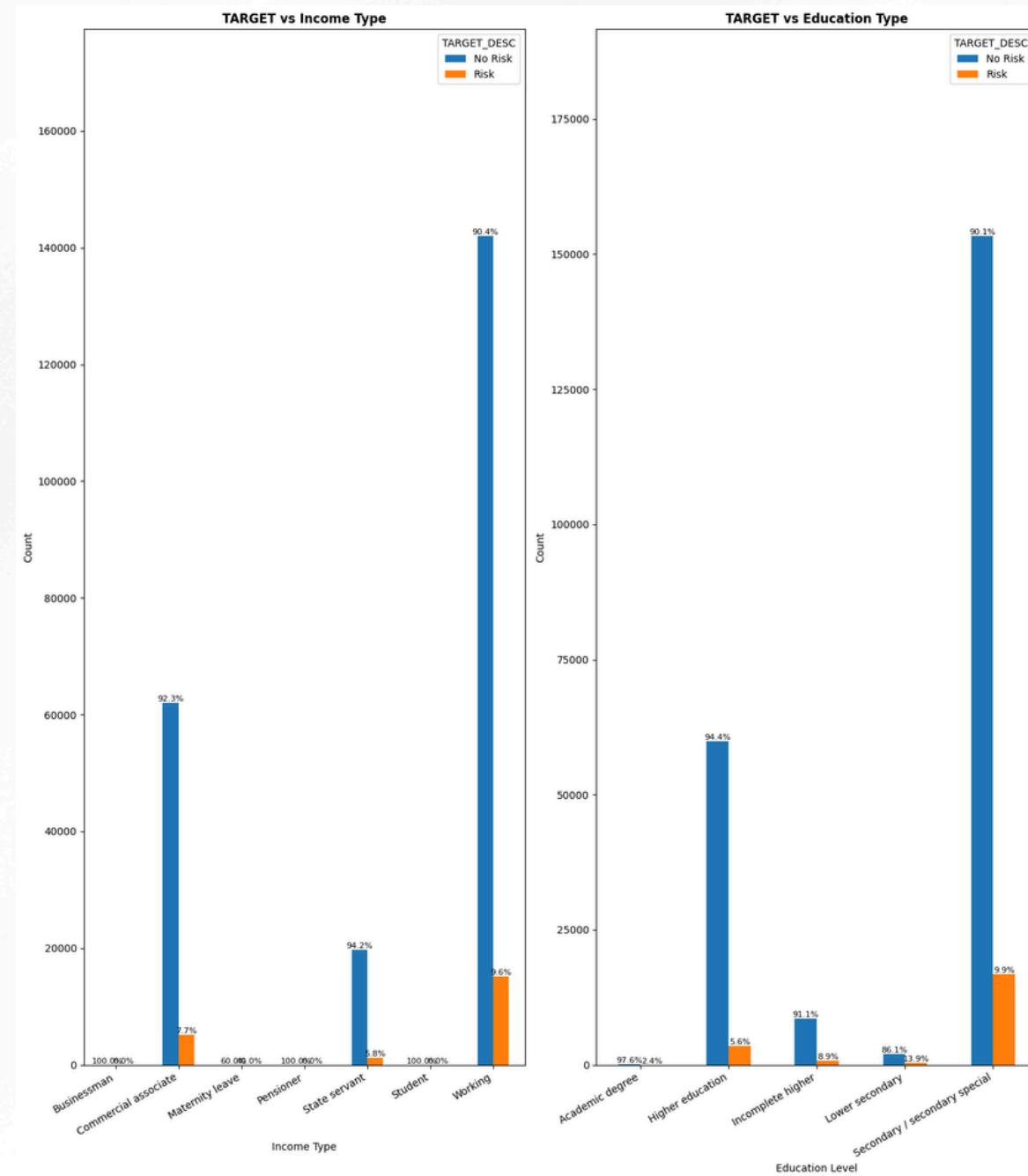
- Young Adults (35 tahun kebawah) memiliki persentase risiko tertinggi (10,9%) dan Older (50 tahun keatas) memiliki tingkat risiko rendah (6,2%).
- Nasabah yang lebih muda cenderung memiliki risiko gagal bayar lebih tinggi
- Pria cenderung lebih berisiko gagal bayar (10,6%) dibanding wanita (7,7%).
- Nasabah single memiliki risiko yang sedikit lebih tinggi (9,7%) dibandingkan nasabah yang sudah menikah (8,5%).
- Cash loans memiliki tingkat risiko lebih tinggi (9,1%) dibandingkan revolving loans (5,8%)

Exploratory Data Analysis



- Nasabah yang tidak memiliki mobil lebih berisiko (9,5%) dibandingkan yang memiliki mobil (7,6%).
- Tidak ada perbedaan signifikan dalam risiko gagal bayar antara nasabah yang memiliki properti atau tidak.

Exploratory Data Analysis



- Maternity leave (40%) dan working (9,6%) memiliki tingkat risiko yang lebih tinggi dibanding yang lain.
- Lower secondary dan Secondary / Secondary special memiliki risiko lebih tinggi dibandingkan Academic degree. Semakin tinggi pendidikan, semakin kecil kemungkinannya untuk gagal bayar.

Feature Engineering



Label Encoder

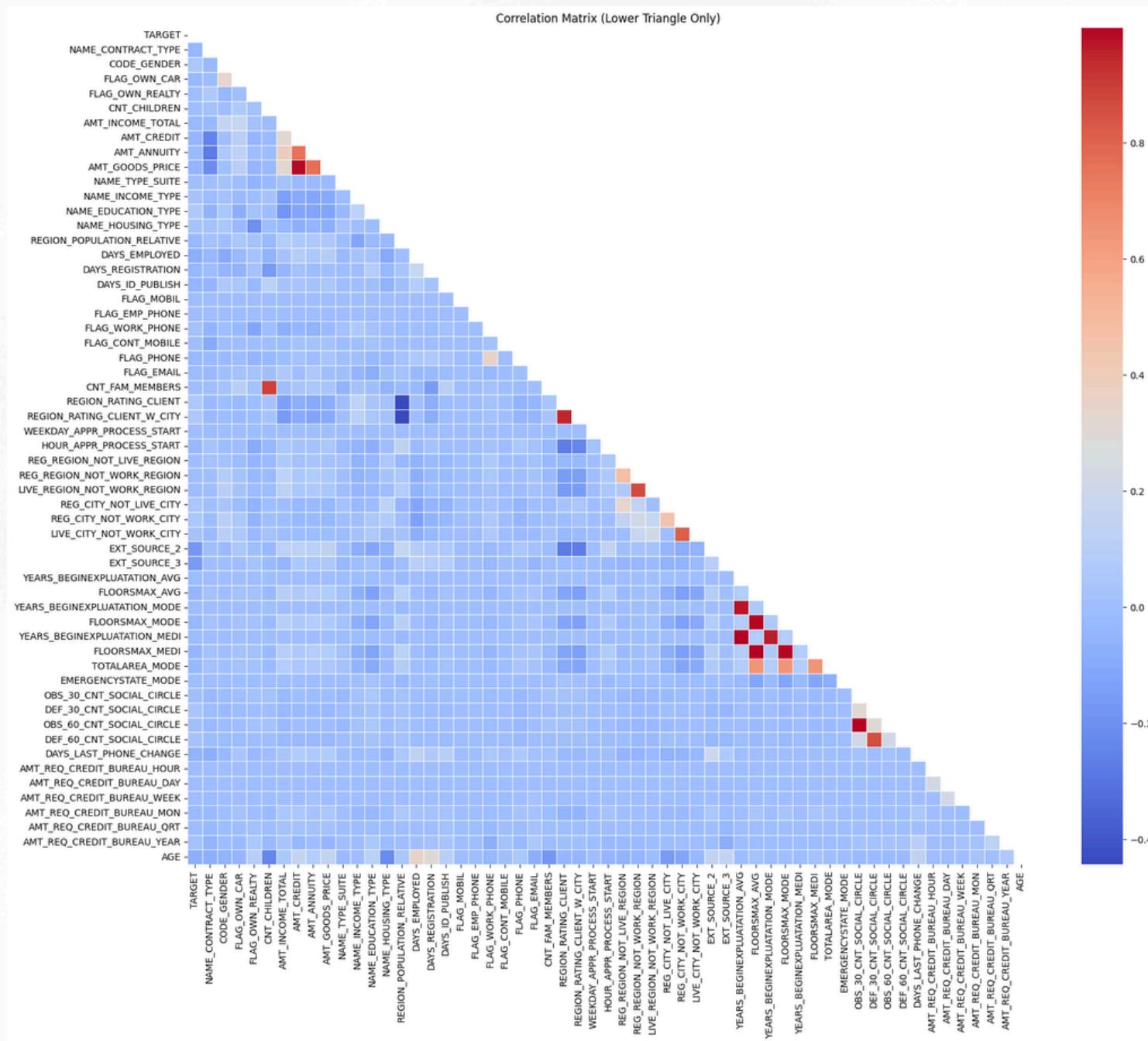
'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY',
'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_HOUSING_TYPE',
'WEEKDAY_APPR_PROCESS_START', 'EMERGENCYSTATE_MODE'

Imbalanced Data

Oversampling data using SMOTE

Feature Selection

Correlation Matrix



Terdapat multikolinearitas yang menunjukkan korelasi kuat antar fitur kolom. Kolom dengan korelasi kuat diatas 0,8 akan dihapus.

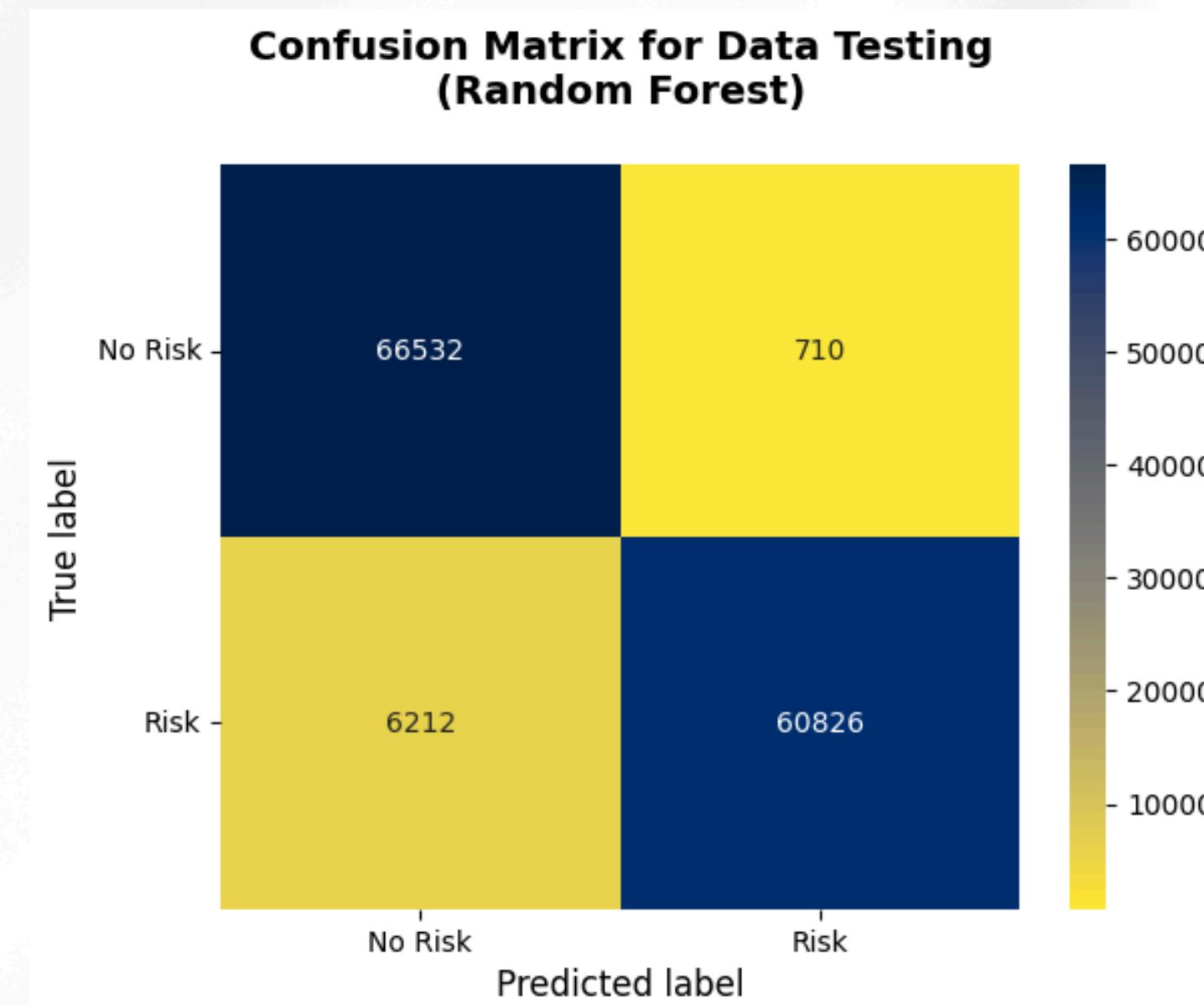
Modelling and Evaluation



Model	Accuracy	Precision	Recall	F1-Score	ROC AUC Score
Logistic Regression	68%	68%	69%	68%	68%
Naives Bayes	62%	63%	58%	61%	62%
K-Nearest Neighbors	82%	73%	100%	85%	81%
Decision Tree	89%	88%	90%	89%	88%
Random Forest	95%	99%	91%	95%	94%

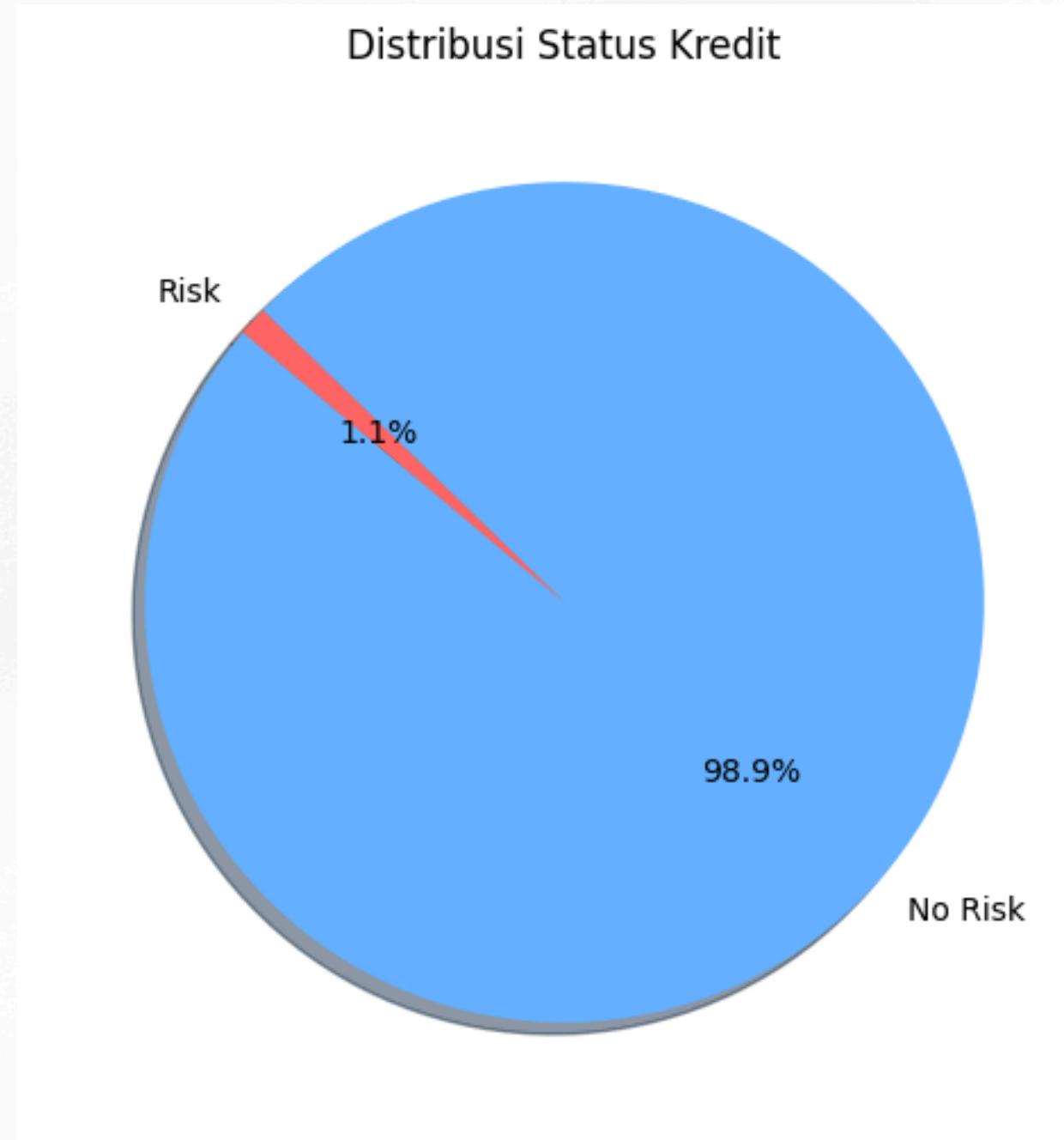
Model Random Forest memiliki tingkat F1-Score dan ROC AUC Score tertinggi.

Modelling and Evaluation



Matriks di atas menunjukkan bahwa dari total 61.536 nasabah yang berada diprediksi berisiko gagal bayar (Risk), model dapat memprediksi 60.826 di antaranya benar-benar berisiko atau sekitar 98,8%.

Prediction Result



TARGET	
0	39014
1	444

- Dataset test yang digunakan pada proyek ini adalah “application_test.csv”
- Dataset terdiri dari 48744 baris dan 121 kolom
- Dengan menggunakan model Random Forest diperoleh bahwa nasabah yang berisiko gagal bayar sebanyak 444 atau 1,1% dari total nasabah.

Conclusion dan Recomendation



Mayoritas nasabah tergolong tidak berisiko.



Pria muda yang belum menikah dan tidak memiliki mobil cenderung lebih berisiko



Random Forest memiliki performa terbaik dengan 95% F1-Score dan 94% ROC Score



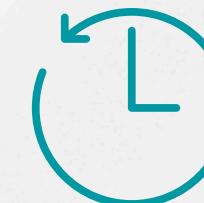
Fokus mitigasi risiko pada faktor yang rawan



Gunakan Random Forest model sebagai baseline



Terapkan strategi kredit berbasis data



Lakukan retraining model secara berkala



Integrasi model ke sistem operasional

Thankyou!



[Yuly Artami Sagala](#)



<https://github.com/yulyartamisagala>