Rakamin Academy X id/x partners

Final Task: Predicting Accepted Loan Scenario using Decision Tree and Random Forest

ID/X Partners Data Scientist VIX

Presented by
Fawwaz Nurmansyah

Fawwaz Nurmansyah

About Me:

A graduate Bachelor of Biology from Airlangga University who is interest in pursuing a career related to data after completing some bootcamp. Able to work well independently and as part of a team, even under tight deadlines and pressure.

Have less than a year experience in data science with understanding of data processing, data analysis, and machine learning. Proficient in extracting insights from complex datasets and transforming them into profitable recommendations for business.

Rakamin Academy

Experience

Final Project Data Science Bootcamp
Rakamin Academy

Doing Exploratory Data Analysis on an E-commerce Shipping Data from Kaggle, preprocessing data, and fitting a Machine Learning model to the dataset using Decision Tree, K-Nearest Neighbour (KNN), Adaboost, Extreme Gradient Boost (XGBoost), Random Forest Classifier, and CATBoost. The best fit models to predict the dataset using XGBoost with highest recall score of 98% and 97% recall cross-validation training and testing data.
Portofolio Link

Case Study

Latar Belakang Tugas

Sebagai tugas akhir dari masa kontrakmu sebagai intern Data Scientist di ID/X Partners, kali ini kamu akan dilibatkan dalam projek dari sebuah lending company. Kamu akan berkolaborasi dengan berbagai departemen lain dalam projek ini untuk menyediakan solusi teknologi bagi company tersebut. Kamu diminta untuk membangun model yang dapat memprediksi credit risk menggunakan dataset yang disediakan oleh company yang terdiri dari data pinjaman yang diterima dan yang ditolak. Selain itu kamu juga perlu mempersiapkan media visual untuk mempresentasikan solusi ke klien. Pastikan media visual yang kamu buat jelas, mudah dibaca, dan komunikatif. Pengerjaan end-to-end solution ini dapat dilakukan di Programming Language pilihanmu dengan tetap mengacu kepada framework/methodology Data Science.

Dataset download here
Data dictionary viewed here

Case Study

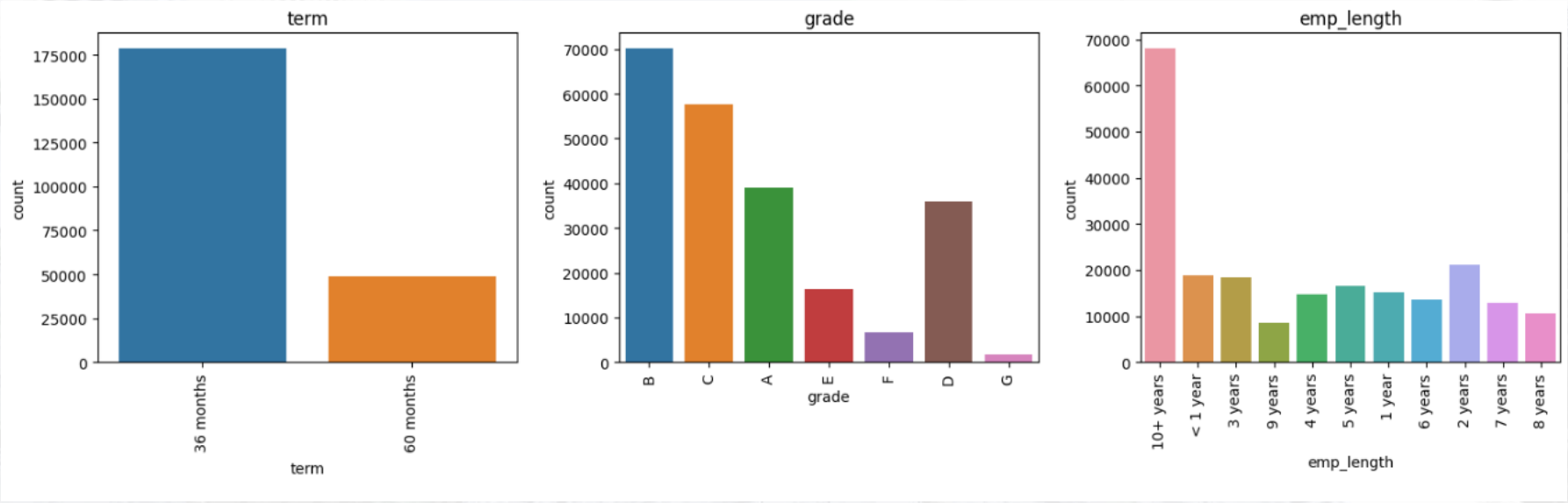
1. Descriptive Analysis

A. Terdapat 17 kolom yang memiliki NULL value keseluruhan pada data yang disediakan.
B. Terdapat 8 kolom yang memiliki high cardinality (terlalu banyak unique value) dan kolom yang tidak diperlukan untuk machine learning.
C. Keseluruhan data yang tersedia di dataset ini adalah sejumlah 466.285 baris/data yang memiliki datatype float64 sejumlah 29 kolom, int64 sejumlah 5 kolom, dan object sejumlah 20 kolom.
D. Kolom yang akan dijadikan target adalah loan_status dengan 9 unique value. Kolom ini akan diambil 2 unique value untuk dijadikan target, yaitu Fully Paid dan Charged Off.
E. Ada beberapa kolom yang memiliki NULL value yang cukup banyak, yaitu next_pymnt_d (49%), mths_since_last_record (43%), mths_since_last_major_derog (40%), mths_since_last_delinq (27%), tot_coll_amt (14%), tot_cur_bal (14%) dan total_rev_hi_lim (14%). Untuk kolom yang memiliki NULL value yang kecil adalah emp_length (1.86%), last_pymnt_d (0.07%), revol_util (0.03%), dan collections_12_mths_ex_med (0.01%).

Case Study

Rakamin Academy

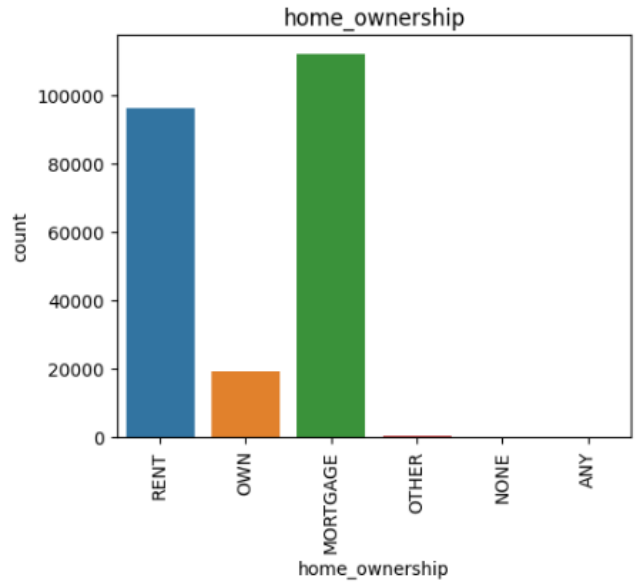
2. Univariate Analysis

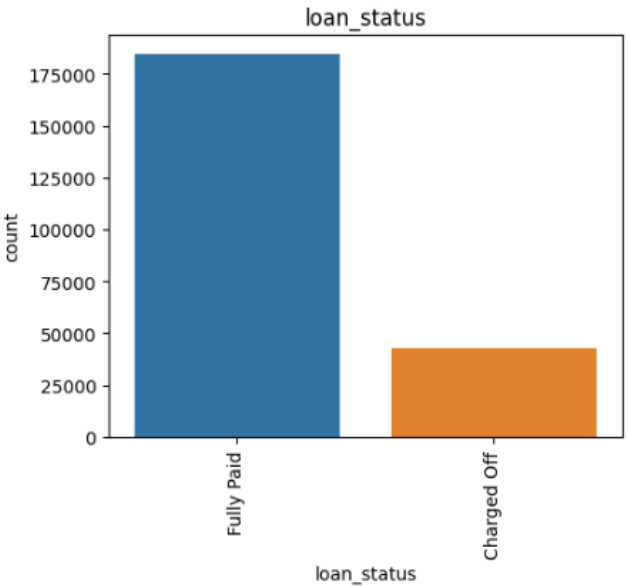


Case Study

Rakamin Academy

2. Univariate Analysis

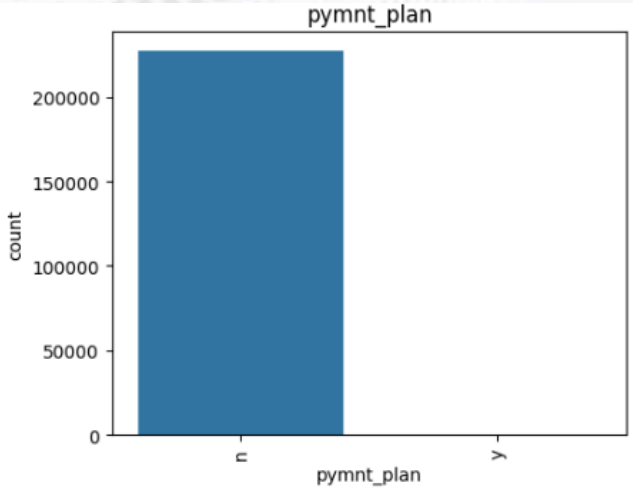


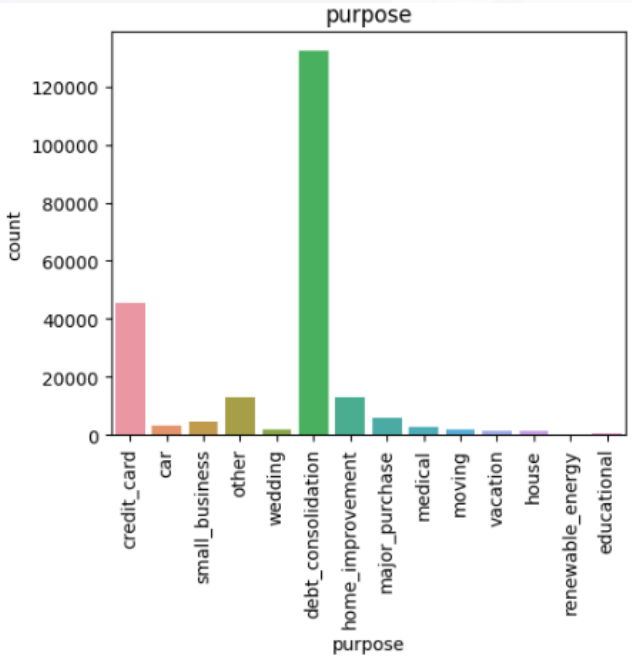


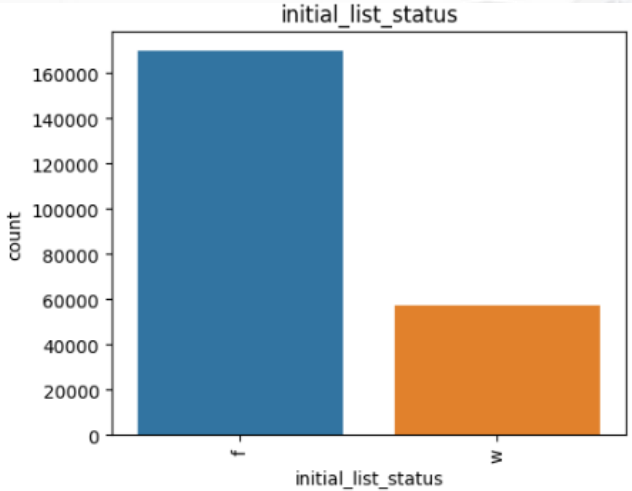
Case Study

Rakamin Academy

2. Univariate Analysis





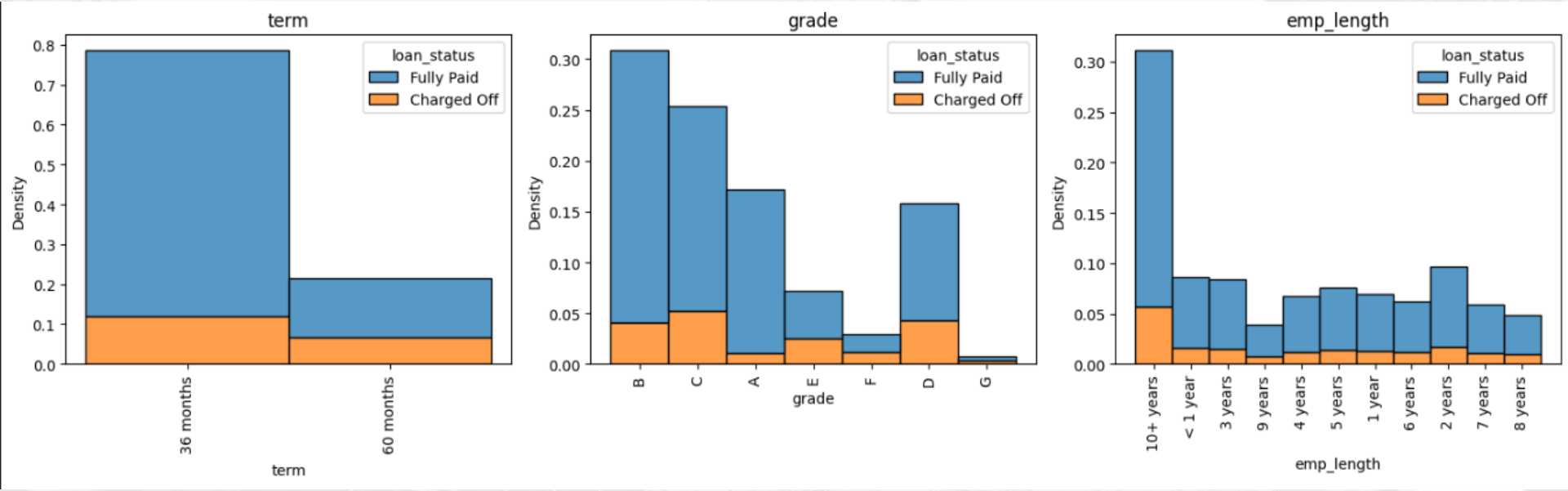


Case Study

Rakamin Academy

2. Univariate Analysis

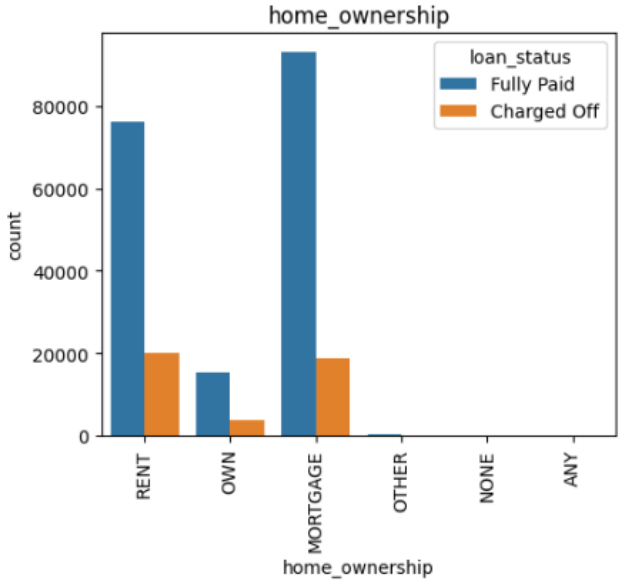
Berdasarkan hasil barplot Univariate Analysis yang ditampilkan, data sudah diambil berdasarkan loan_status yang memiliki value ‘Fully Paid’ dan ‘Charged Off’.

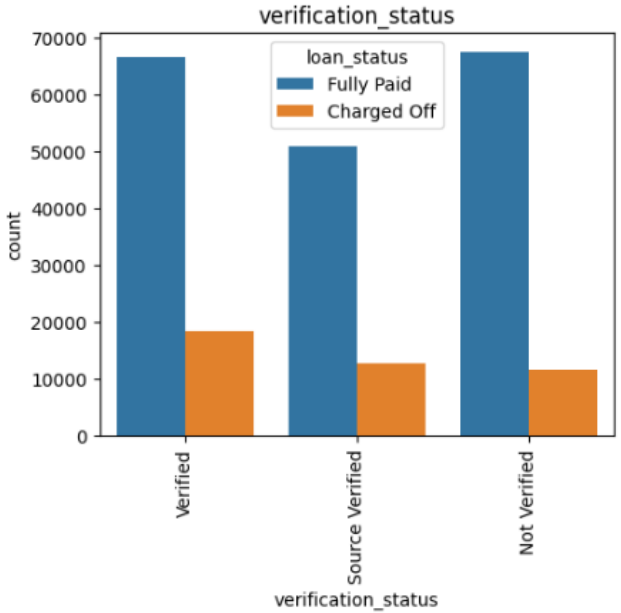


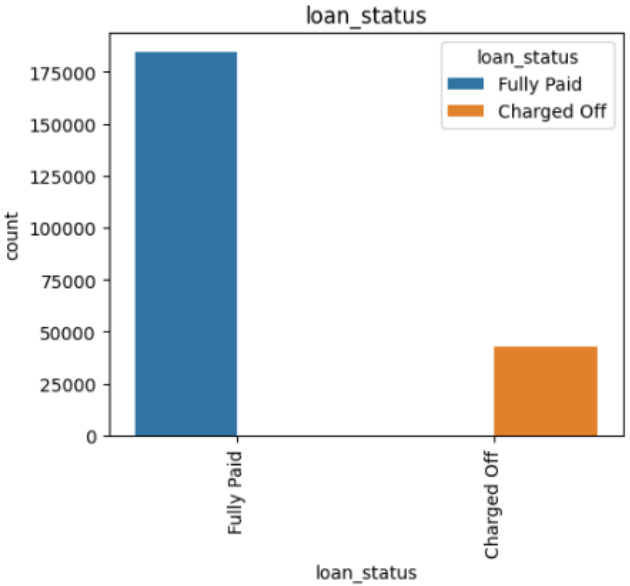
Case Study

Rakamin Academy

2. Univariate Analysis



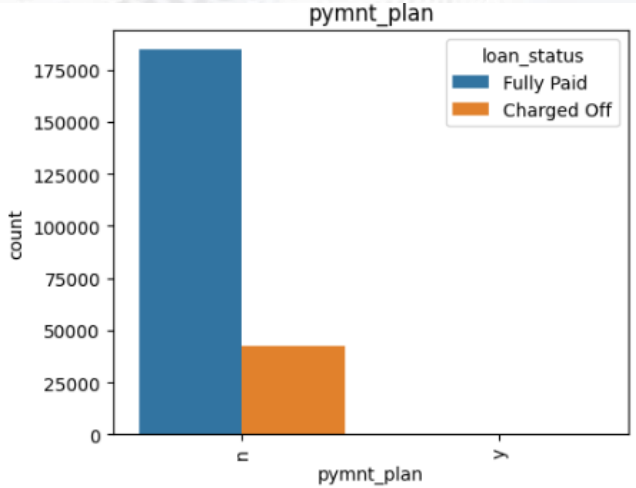


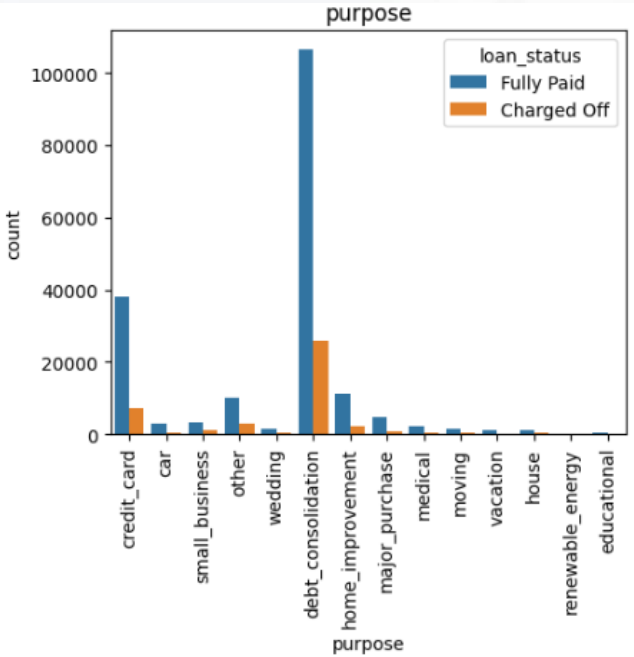


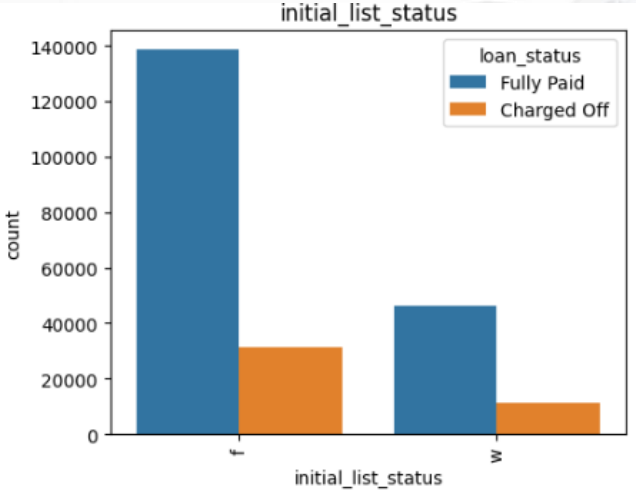
Case Study

Rakamin Academy

2. Univariate Analysis







Case Study

Rakamin Academy

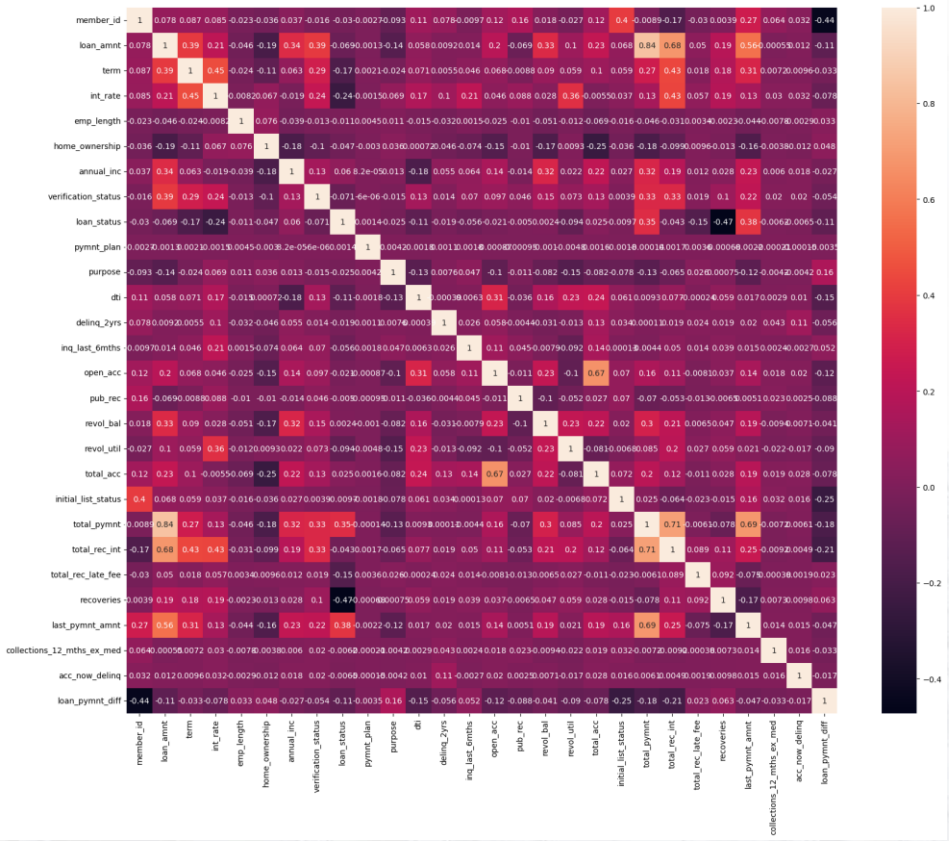
2. Univariate Analysis

Pada histplot yang disediakan di .ipynb, tidak ada signifikansi bentuk data. Semua data merupakan data yang negatively skewed.

Case Study

Rakamin Academy

3. Multivariate Analysis



Case Study

Rakamin Academy

4. Data Preprocessing

A. Sebelum EDA,
- Drop kolom yang memiliki semua valuenya NULL.
- Drop kolom yang tidak dibutuhkan untuk machine learning
- Drop kolom yang memiliki high cardinality value (memiliki banyak unique value)
- Semua tanggal, bulan, dan tahun yang semulanya berupa datatype object, diubah menjadi datetime.

A. Setelah EDA,
- Drop kolom yang memiliki missing value yang lebih dari 10%
- Drop data yang memiliki missing value
- Drop fitur yang tidak dibutuhkan untuk machine learning (sub_grade)
- Tidak ada nilai yang duplikat
- Ekstraksi fitur last_pymnt_d menjadi loan_pymnt_diff untuk menghitung berapa hari semenjak terakhir membayar loan.
- Melakukan feature scaling untuk numeric data dan feature encoding untuk categorical data

Case Study

Rakamin Academy

4. Data Preprocessing

A. Sebelum EDA,
- Drop kolom yang memiliki semua valuenya NULL.
- Drop kolom yang tidak dibutuhkan untuk machine learning
- Drop kolom yang memiliki high cardinality value (memiliki banyak unique value)
- Semua tanggal, bulan, dan tahun yang semulanya berupa datatype object, diubah menjadi datetime.

A. Setelah EDA,
- Drop kolom yang memiliki missing value yang lebih dari 10%
- Drop data yang memiliki missing value
- Drop fitur yang tidak dibutuhkan untuk machine learning (sub_grade)
- Tidak ada nilai yang duplikat
- Ekstraksi fitur last_pymnt_d menjadi loan_pymnt_diff untuk menghitung berapa hari semenjak terakhir membayar loan.
- Melakukan feature scaling untuk numeric data dan feature encoding untuk categorical data

Case Study

Rakamin Academy

5. Modelling

```
Original (y Train)
loan_status
1    142548
0     31828
Name: count, dtype: int64


y SMOTE
loan_status
0    39768
1    39768
Name: count, dtype: int64


Original (X Train)
(174376, 27)


X SMOTE
(79536, 27)
```

Sebelum melakukan modelling, dilakukan SMOTE untuk menyeimbangkan class yang akan dipelajari oleh machine learning.
Data yang dipelajari sangat banyak sehingga memutuskan untuk melakukan Undersampling.

Case Study

Rakamin Academy

5. Modelling

Modelling dilakukan dengan menggunakan Random Forest dan Decision Tree dengan dua perlakukan, yaitu tanpa SMOTE dan dengan SMOTE.

Random Forest

Random Forest

Without SMOTE

```
from sklearn.ensemble import RandomForestClassifier# import decision tree dari sklearn

rfc = RandomForestClassifier(random_state=727) # inisiasi object dengan nama rfc
rfc.fit(X_train, y_train)
eval_classification(rfc)
eval_cv_ab_roc_auc(rfc)

# Show feature importance
# show_feature_importance(rfc.best_estimator_)
```

✓ 4m 13.0s

```
Accuracy (Test Set): 1.00
Precision (Test Set): 0.99
Recall (Test Set): 1.00
F1-Score (Test Set): 1.00

roc_auc (train-proba): 1.00
roc_auc (test-proba): 1.00
roc_auc (crossval train): 1.0
roc_auc (crossval test): 0.9994579496956078
```

With SMOTE

```
from sklearn.ensemble import RandomForestClassifier# import decision tree dari sklearn

rfc = RandomForestClassifier(random_state=727) # inisiasi object dengan nama dt
rfc.fit(X_under_SMOTE, y_under_SMOTE)
eval_classification2(rfc)
eval_cv_ab_roc_auc2(rfc)

# Show feature importance
imp_df = pd.DataFrame({
    "Feature Name": X_under_SMOTE.columns,
    "Importance": rfc.feature_importances_
})
fi = imp_df.sort_values(by="Importance", ascending=False)
fi2 = fi.head(10)
plt.figure(figsize=(10,8))
sns.barplot(data=fi2, x='Importance', y='Feature Name')
plt.title('Top 10 Feature Importance Each Attributes (Random Forest)', fontsize=18)
plt.xlabel ('Importance', fontsize=16)
plt.ylabel ('Feature Name', fontsize=16)
plt.show()
```

✓ 1m 36.0s

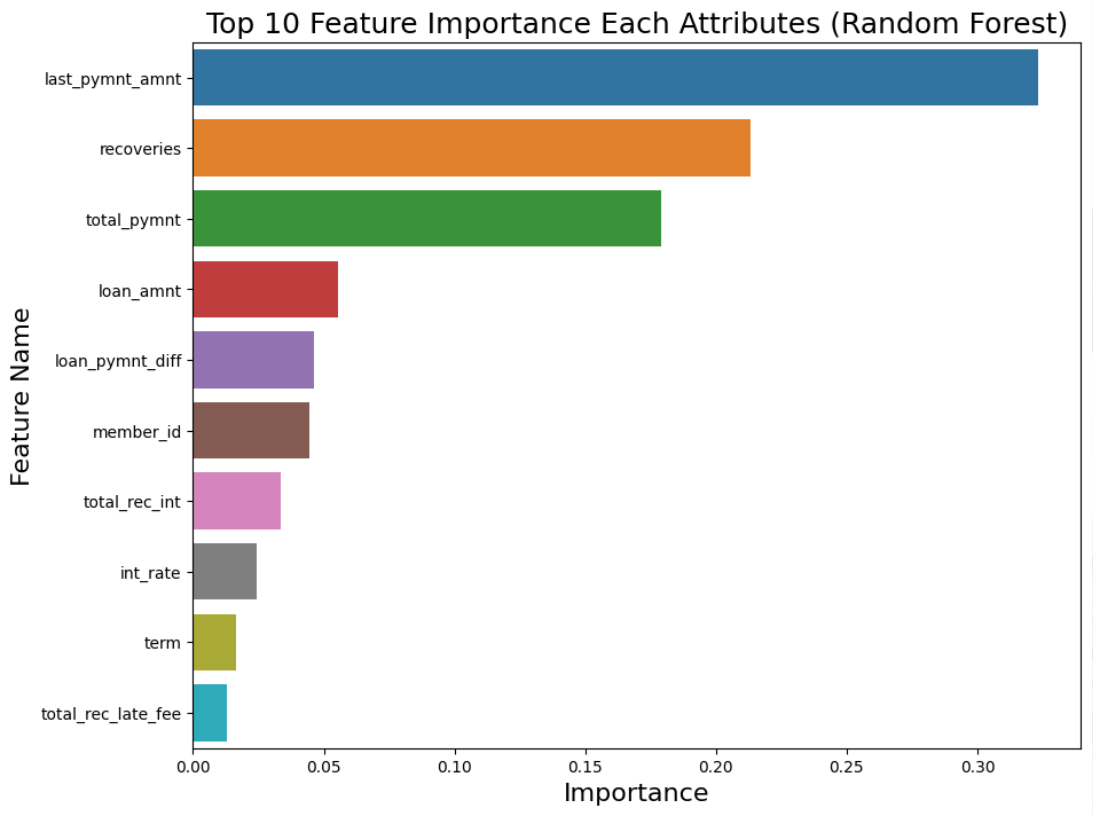
```
Accuracy (Test Set): 1.00
Precision (Test Set): 1.00
Recall (Test Set): 1.00
F1-Score (Test Set): 1.00

roc_auc (train-proba): 1.00
roc_auc (test-proba): 1.00
roc_auc (crossval train): 1.0
roc_auc (crossval test): 0.9969753276764017
```

Case Study

Rakamin Academy

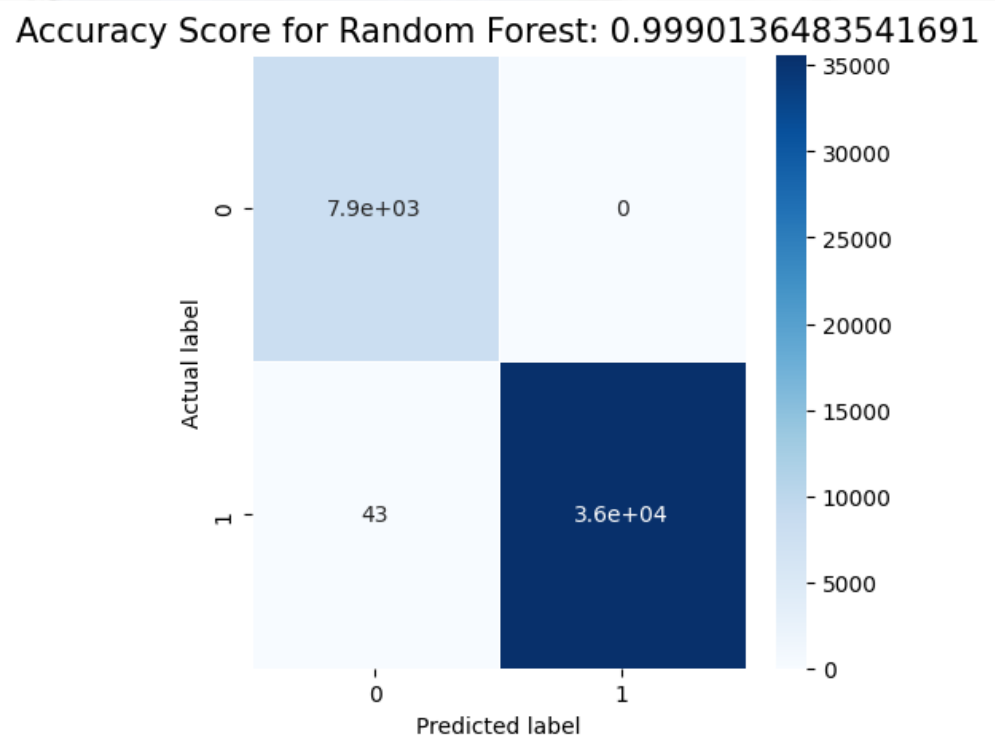
5. Modelling



Case Study

Rakamin Academy

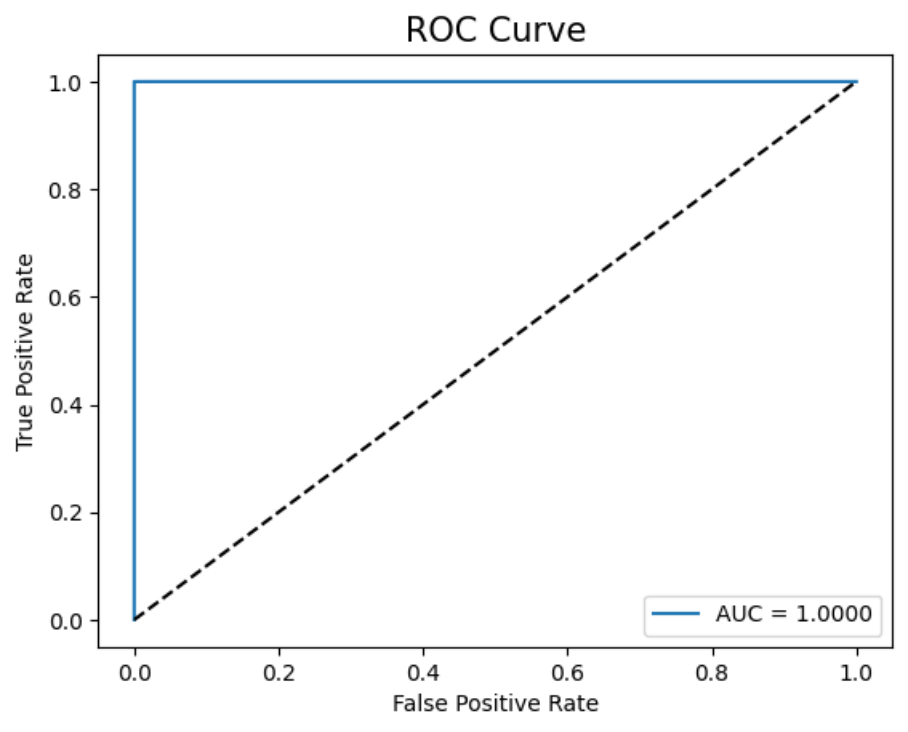
5. Modelling



Case Study

Rakamin Academy

5. Modelling



Case Study

Rakamin Academy

5. Modelling

Decision Tree

Without SMOTE

```
from sklearn.tree import DecisionTreeClassifier# import decision tree dari sklearn

dt = DecisionTreeClassifier(random_state=727) # inisiasi object dengan nama dt
dt.fit(X_train, y_train)
eval_classification(dt)
eval_cv_ab_roc_auc(dt)

# Show feature importance
# show_feature_importance(dt.best_estimator_)
```

✓ 18.4s

```
Accuracy (Test Set): 0.99
Precision (Test Set): 1.00
Recall (Test Set): 1.00
F1-Score (Test Set): 1.00

roc_auc (train-proba): 1.00
roc_auc (test-proba): 0.99
roc_auc (crossval train): 1.0
roc_auc (crossval test): 0.9888396806097001
```

With SMOTE

```
from sklearn.tree import DecisionTreeClassifier# import decision tree dari sklearn

dt = DecisionTreeClassifier(random_state=727) # inisiasi object dengan nama dt
dt.fit(X_under_SMOTE, y_under_SMOTE)
eval_classification2(dt)
eval_cv_ab_roc_auc2(dt)

# Show feature importance
imp_df = pd.DataFrame({
    "Feature Name": X_under_SMOTE.columns,
    "Importance": dt.feature_importances_
})
fi = imp_df.sort_values(by="Importance", ascending=False)
fi2 = fi.head(10)
plt.figure(figsize=(10,8))
sns.barplot(data=fi2, x='Importance', y='Feature Name')
plt.title('Top 10 Feature Importance Each Attributes (Decision Tree)', fontsize=18)
plt.xlabel ('Importance', fontsize=16)
plt.ylabel ('Feature Name', fontsize=16)
plt.show()
# Show feature importance
# show_feature_importance(dt.best_estimator_)
```

[61] ✓ 7.2s

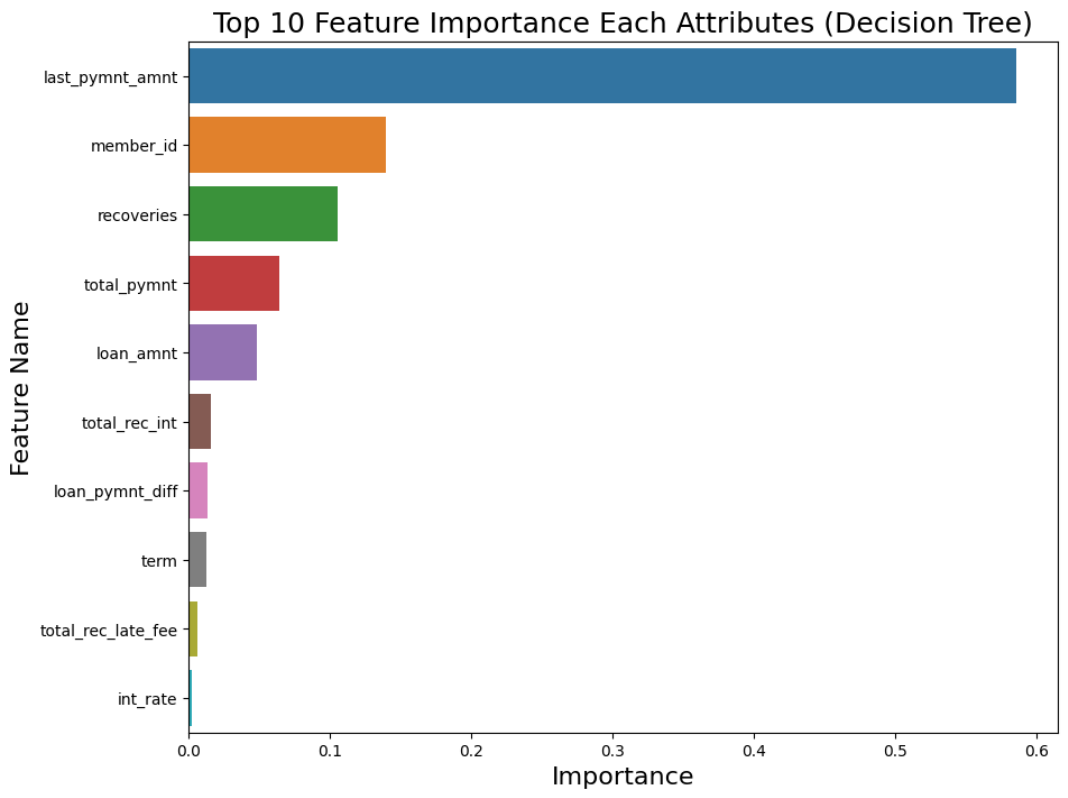
```
Accuracy (Test Set): 0.99
Precision (Test Set): 1.00
Recall (Test Set): 0.99
F1-Score (Test Set): 1.00

roc_auc (train-proba): 1.00
roc_auc (test-proba): 1.00
roc_auc (crossval train): 1.0
roc_auc (crossval test): 0.9168157525664435
```

Case Study

Rakamin Academy

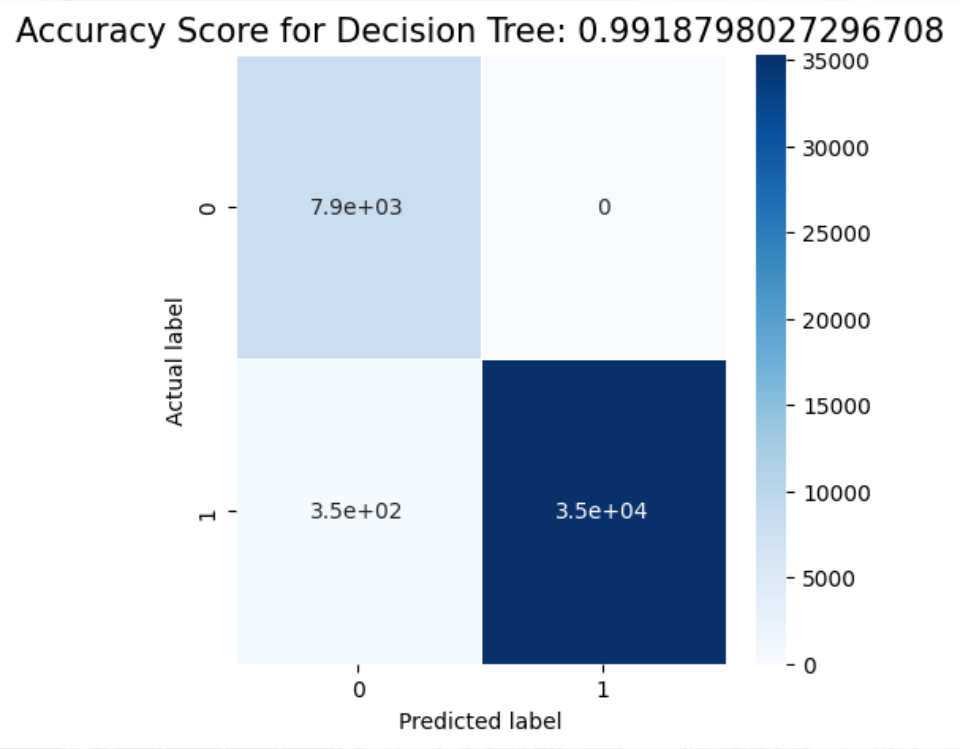
5. Modelling



Case Study

Rakamin Academy

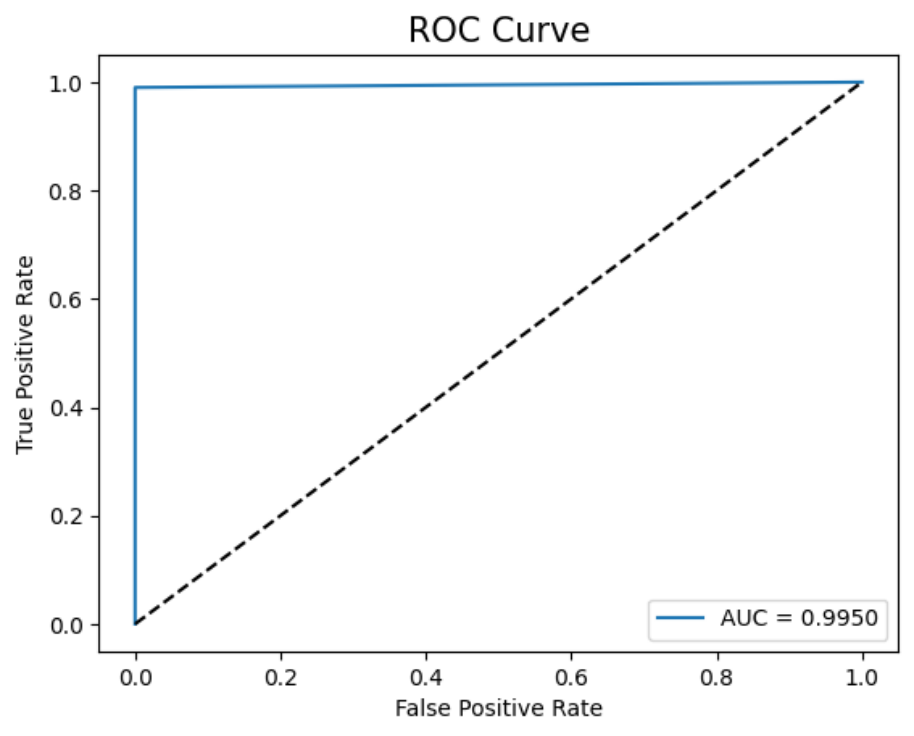
5. Modelling



Case Study

Rakamin Academy

5. Modelling



Rakamin Academy

Link Github

https://github.com/FawwazN/data-scientist-idx

Thank You