Asmiyeni Islamiati asmiyeniislamiati@gmail.com

Developing a Predictive Model for Credit Risk Assessment

Project Based Internship Data Scientist ID/X Partner

Company Overview

id/x partners was established in 2002 by ex-bankers and management consultants with extensive experience in credit cycle and process management, scoring development, and performance management. Their combined expertise has served corporations across Asia and Australia in multiple industries, specifically financial services, telecommunications, manufacturing, and retail.

id/x partners provides consulting services specializing in utilizing data analytics and decisioning (DAD) solutions combined with an integrated risk management and marketing discipline to help clients optimize portfolio profitability and business processes. Comprehensive consulting service and technology solutions offered by id/x partners position them as a one-stop service provider.

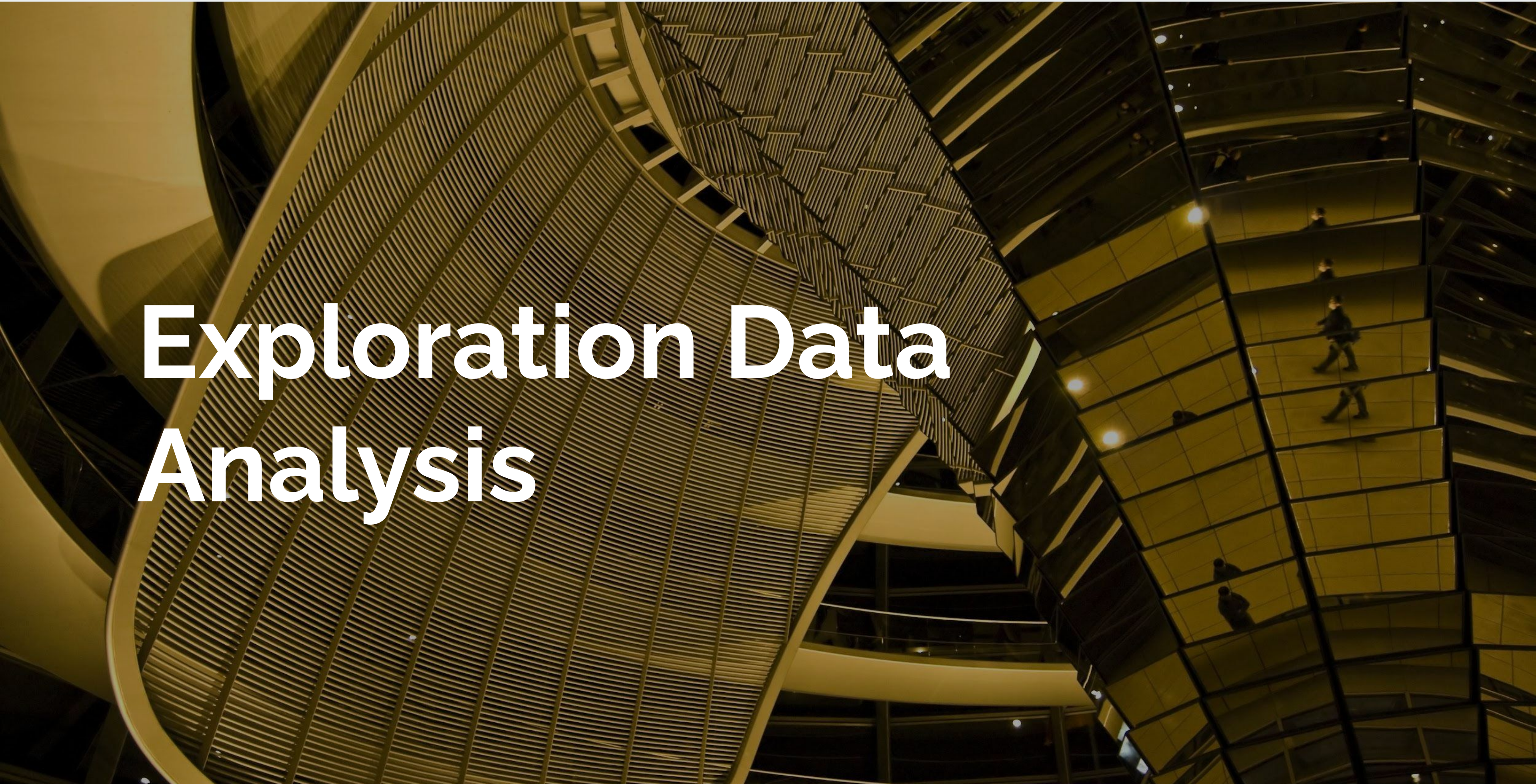
Project objective

As a Data Scientist at ID/X Partners, I am involved in a project for a lending company (multifinance) aimed at improving the accuracy in assessing and managing credit risk. This project involves developing a machine learning model using approved and rejected loan datasets to predict credit risk.

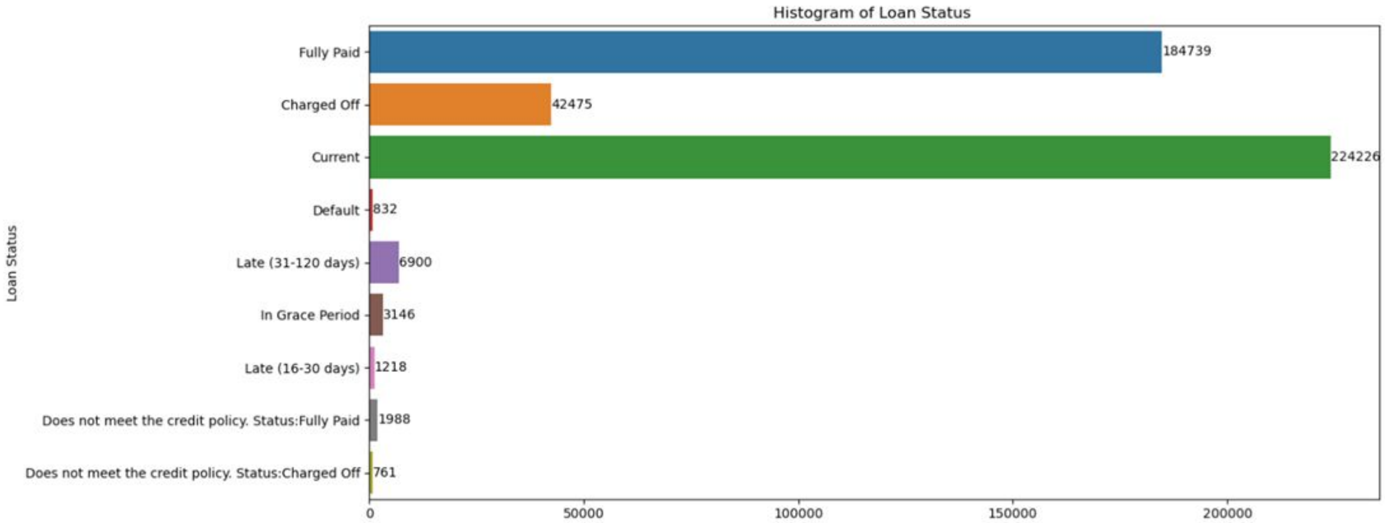
Objective

1. Conduct Exploratory Data Analysis (EDA) to understand the dataset and identify patterns.
2. Perform data preprocessing to clean and prepare the data for modeling.
3. Develop machine learning models to predict credit risk.
4. Evaluate the performance of the machine learning models to ensure their accuracy and reliability.

Exploration Data Analysis



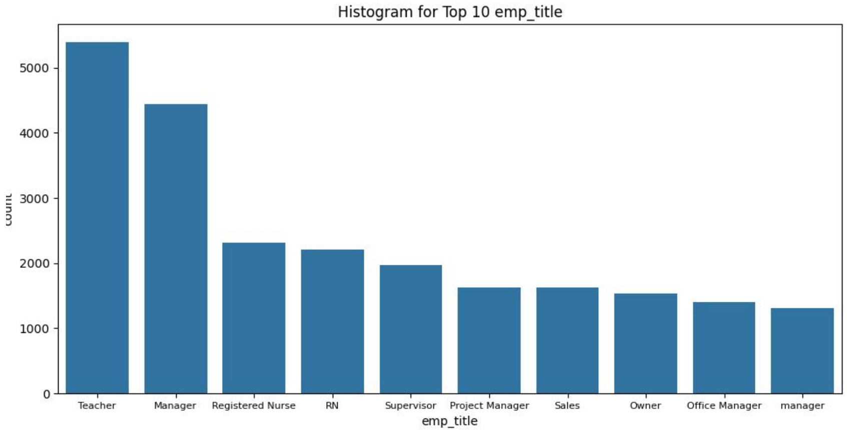
Market trends

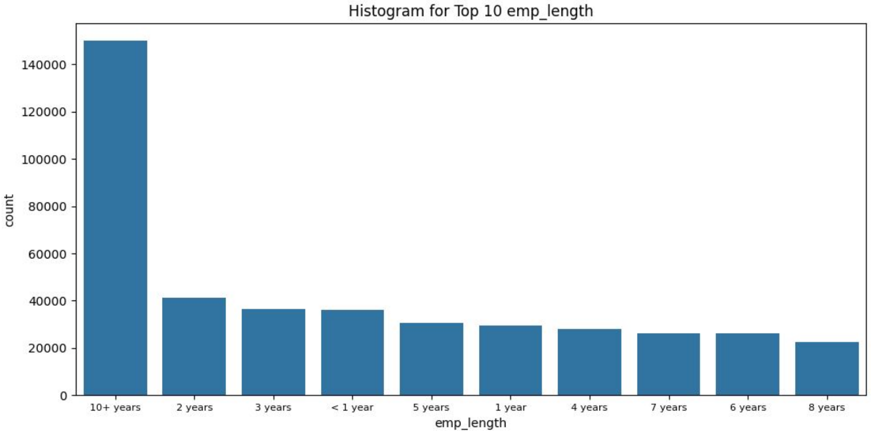


The 'loan_status' column contains more than two unique values and will be divided into 'good_loan' for loans considered safe and 'bad_loan' for those deemed risky, with the primary goal of achieving accuracy and efficiency in credit risk management.

The 'emp_title' column represents the job title provided by the borrower when applying for a loan. The order of visualization is Teacher, Manager, and Registered Nurse.

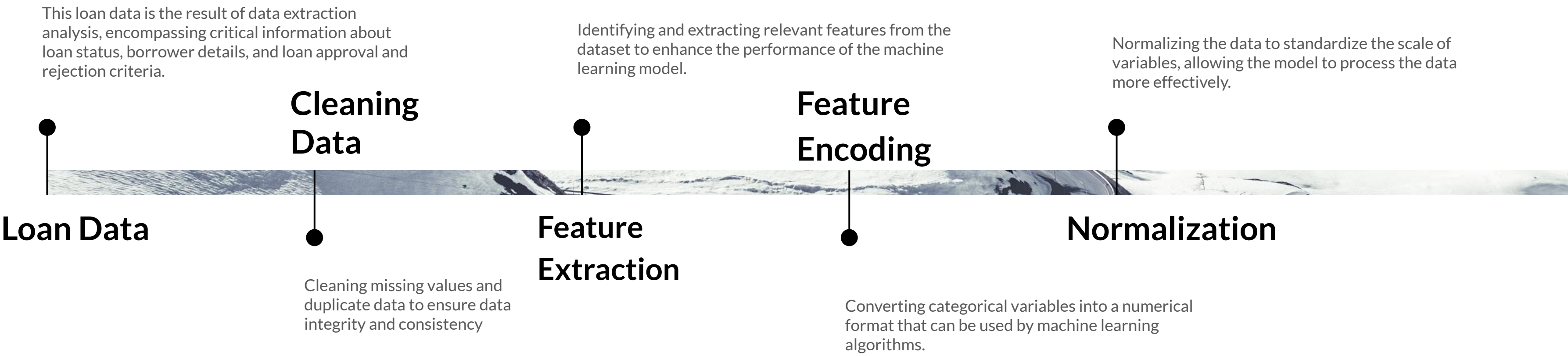
The 'emp_length' column indicates the length of employment in years. The visualization order is as follows: '10+ years' leads with over 140,000 borrowers, followed by '2 years' with 40,000 borrowers. The numbers gradually decrease for the remaining employment lengths, but the differences are not significant.





Data Preprocessing

Diagram PreProcessing



Loan Data

This loan data is the result of data extraction analysis, encompassing critical information about loan status, borrower details, and loan approval and rejection criteria.

Cleaning Data

Cleaning missing values and duplicate data to ensure data integrity and consistency

Feature Extraction

Identifying and extracting relevant features from the dataset to enhance the performance of the machine learning model.

Feature Encoding

Converting categorical variables into a numerical format that can be used by machine learning algorithms.

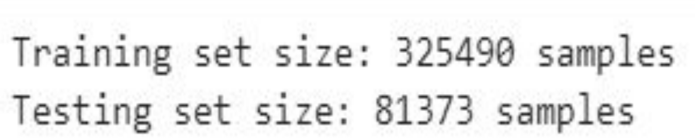
Normalization

Normalizing the data to standardize the scale of variables, allowing the model to process the data more effectively.

Modelling

Train Model

Split Train-Test Model



```
Training set size: 325490 samples
Testing set size: 81373 samples
```

Algorithm Modelling

01 | Logistic Regression = `logreg_model`

02 | Random Forest = `dt_model`

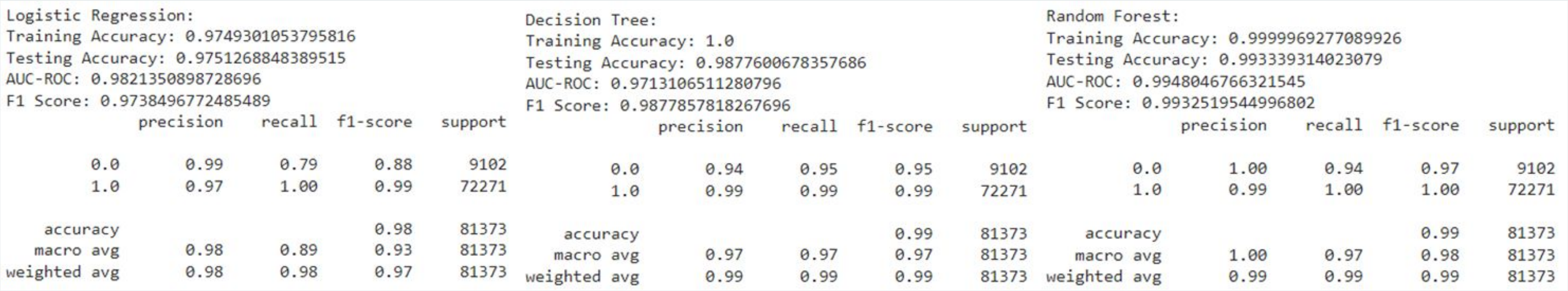
03 | Decision Tree = `rf_model`

1. Train test split is a technique used to partition a dataset into training and testing subsets, allowing the model to be trained on the training set and evaluated on the unseen testing set to assess its generalization ability.

2. The selection of Logistic Regression, Decision Tree, and Random Forest models is based on their suitability for classification tasks like predicting credit risk, offering varying complexities and performance metrics to explore and compare across different types of machine learning algorithms.

Model Evaluation

Before Hyperparameter Tuning

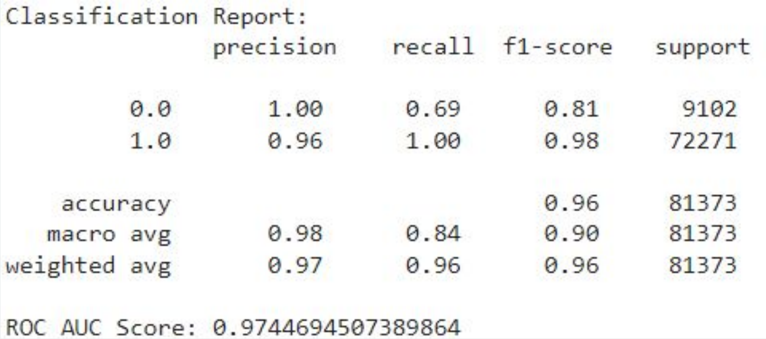


```
Logistic Regression:
Training Accuracy: 0.9749301053795816
Testing Accuracy: 0.9751268848389515
AUC-ROC: 0.9821350898728696
F1 Score: 0.9738496772485489
              precision    recall  f1-score   support

         0.0       0.99      0.79      0.88      9102
         1.0       0.97      1.00      0.99     72271

    accuracy                           0.98     81373
   macro avg       0.98      0.89      0.93     81373
weighted avg       0.98      0.98      0.97     81373
```

```
Decision Tree:
Training Accuracy: 1.0
Testing Accuracy: 0.9877600678357686
AUC-ROC: 0.9713106511280796
F1 Score: 0.9877857818267696
              precision    recall  f1-score   support

         0.0       0.94      0.95      0.95      9102
         1.0       0.99      0.99      0.99     72271

    accuracy                           0.99     81373
   macro avg       0.97      0.97      0.97     81373
weighted avg       0.99      0.99      0.99     81373
```

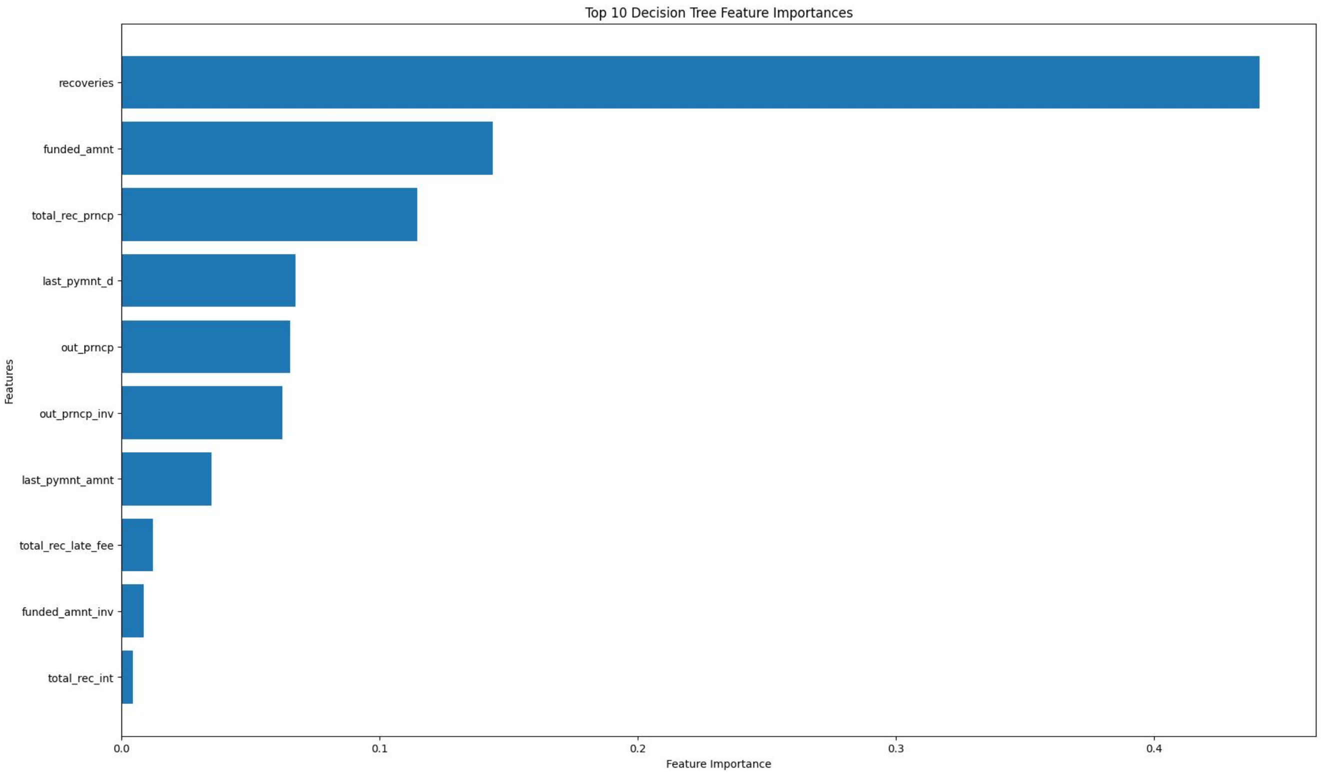
```
Random Forest:
Training Accuracy: 0.9999969277089926
Testing Accuracy: 0.993339314023079
AUC-ROC: 0.9948046766321545
F1 Score: 0.9932519544996802
              precision    recall  f1-score   support

         0.0       1.00      0.94      0.97      9102
         1.0       0.99      1.00      1.00     72271

    accuracy                           0.99     81373
   macro avg       1.00      0.97      0.98     81373
weighted avg       0.99      0.99      0.99     81373
```

After Hyperparameter Tuning



```
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      0.69      0.81      9102
         1.0       0.96      1.00      0.98     72271

    accuracy                           0.96     81373
   macro avg       0.98      0.84      0.90     81373
weighted avg       0.97      0.96      0.96     81373

ROC AUC Score: 0.9744694507389864
```

Feature Importance



Definisi Feature

- **recoveries** : Menunjukkan apakah rencana pembayaran telah diterapkan untuk pinjaman
- **funded_amnt** : Jumlah total yang berkomitmen untuk pinjaman tersebut pada saat itu.
- **total_rec_prncp** : Kepala Sekolah diterima sampai saat ini
- **last_pymnt_d** : Pembayaran bulan lalu telah diterima
- **out_prncp** : Sisa pokok terutang untuk jumlah total yang didanai
- **out_prncp_inv** : Sisa pokok terutang untuk sebagian dari jumlah total yang didanai oleh investor
- **last_pymnt_amnt** : Jumlah total pembayaran terakhir yang diterima
- **total_rec_late_fee** : Biaya keterlambatan diterima sampai saat ini
- **funded_amnt_inv** : Jumlah total yang berkomitmen untuk pinjaman tersebut pada saat itu
- **total_rec_int** : Bunga yang diterima sampai saat ini

Business Insight

Conclution

Setelah melakukan evaluasi terhadap beberapa model machine learning, termasuk Logistic Regression, Decision Tree, dan Random Forest, hasil menunjukkan bahwa model Random Forest memberikan hasil pemodelan terbaik untuk memprediksi risiko kredit. Hal tersebut karena:

1.Model Random Forest menghasilkan akurasi yang lebih tinggi dibandingkan dengan model lain. Hal ini menunjukkan bahwa model ini lebih mampu mengklasifikasikan status pinjaman dengan benar.

2.AUC (Area Under the Curve) score untuk Random Forest lebih tinggi, menunjukkan bahwa model ini memiliki kemampuan yang lebih baik dalam membedakan antara kategori 'good_loan' dan 'bad_loan’.

3.Random Forest terkenal dengan kemampuannya yang robust terhadap overfitting, terutama ketika dibandingkan dengan Decision Tree. Model ini mampu menangani dataset yang besar dan kompleks dengan lebih baik.

4.Random Forest memberikan informasi yang berguna mengenai pentingnya setiap fitur dalam prediksi, memungkinkan kita untuk memahami faktor-faktor utama yang mempengaruhi risiko kredit.

Business Recomendation

After evaluating several machine learning models, including Logistic Regression, Decision Tree, and Random Forest, the results indicated that the Random Forest model provides the best performance for predicting credit risk. This is because:

1. The Random Forest model produced higher accuracy compared to other models, demonstrating its superior ability to correctly classify loan statuses.

2. The AUC (Area Under the Curve) score for Random Forest was higher, indicating that this model has a better capability to distinguish between 'good_loan' and 'bad_loan' categories.

3. Random Forest is known for its robustness against overfitting, especially when compared to Decision Trees. It handles large and complex datasets more effectively.

4. Random Forest provides valuable information about the importance of each feature in the prediction, allowing us to understand the key factors influencing credit risk.

Thank you.