FINAL PROJECT RAKAMIN ACADEMY

Credit Risk Prediction Model

Laode Alif Ma'sum Sidrajat Raja Ika

Biodata

- Nama : Laode Alif Ma'sum Sidrajat Raja Ika
- NPM : 2106731213
- Asal Universitas : Universitas Indonesia
- Program Studi : Teknik Komputer

Link Github & Video

Github

Youtube

Cara mempersiapkan ruang kelas virtual

1 — 2 — 3 — 4 — 5

Data Understanding

Data exploration and basic analysis of the dataset's component

Exploratory Data Analysis

Visualization of correlation between components of dataset

Data Preparation

Data preprocessing before using it in modelling. (Normalization, Encoding, Removal, Splitting, etc)

Data Modelling

Machine learning modelling on cleaned dataset (training & testing)

Evaluation

Analyze Evaluate the result of modelling

Data Understanding

```
# Load the datasets from CSV files
loan_df = pd.read_csv("/content/drive/MyDrive/Dataset/loan_data_2007_2014.csv")
pd.set_option('display.max_columns', None)
print(loan_df.head())
```

```
   Unnamed: 0       id  member_id  loan_amnt  funded_amnt  funded_amnt_inv  \
0           0  1077501    1296599       5000         5000           4975.0
1           1  1077430    1314167       2500         2500           2500.0
2           2  1077175    1313524       2400         2400           2400.0
3           3  1076863    1277178      10000        10000          10000.0
4           4  1075358    1311748       3000         3000           3000.0

         term  int_rate  installment grade sub_grade  \
0   36 months     10.65       162.87     B        B2
1   60 months     15.27        59.83     C        C4
2   36 months     15.96        84.33     C        C5
3   36 months     13.49       339.31     C        C1
4   60 months     12.69        67.79     B        B5

                 emp_title emp_length home_ownership  annual_inc  \
0                      NaN  10+ years           RENT     24000.0
1                   Ryder    < 1 year           RENT     30000.0
2                      NaN  10+ years           RENT     12252.0
3       AIR RESOURCES BOARD  10+ years           RENT     49200.0
4  University Medical Group     1 year           RENT     80000.0

  verification_status issue_d  loan_status pymnt_plan  \
0            Verified  Dec-11   Fully Paid          n
1     Source Verified  Dec-11  Charged Off          n
2        Not Verified  Dec-11   Fully Paid          n
3     Source Verified  Dec-11   Fully Paid          n
4     Source Verified  Dec-11      Current          n

                                                 url  \
0  https://www.lendingclub.com/browse/loanDetail....
1  https://www.lendingclub.com/browse/loanDetail....
2  https://www.lendingclub.com/browse/loanDetail....
3  https://www.lendingclub.com/browse/loanDetail....
4  https://www.lendingclub.com/browse/loanDetail....
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		id	member_i	loan_amn	funded_ar	funded_ar	term	int_rate	installmer	grade	sub_grade	emp_title	emp_leng	home_ow
2	0	1077501	1296599	5000	5000	4975	36 month	10.65	162.87	B	B2		10+ years	RENT
3	1	1077430	1314167	2500	2500	2500	60 month	15.27	59.83	C	C4	Ryder	< 1 year	RENT
4	2	1077175	1313524	2400	2400	2400	36 month	15.96	84.33	C	C5		10+ years	RENT
5	3	1076863	1277178	10000	10000	10000	36 month	13.49	339.31	C	C1	AIR RESOU	10+ years	RENT
6	4	1075358	1311748	3000	3000	3000	60 month	12.69	67.79	B	B5	University	1 year	RENT
7	5	1075269	1311441	5000	5000	5000	36 month	7.9	156.46	A	A4	Veolia Tra	3 years	RENT
8	6	1069639	1304742	7000	7000	7000	60 month	15.96	170.08	C	C5	Southern S	8 years	RENT
9	7	1072053	1288686	3000	3000	3000	36 month	18.64	109.43	E	E1	MKC Acco	9 years	RENT
10	8	1071795	1306957	5600	5600	5600	60 month	21.28	152.39	F	F2		4 years	OWN
11	9	1071570	1306721	5375	5375	5350	60 month	12.69	121.45	B	B5	Starbucks	< 1 year	RENT
12	10	1070078	1305201	6500	6500	6500	60 month	14.65	153.45	C	C3	Southwest	5 years	OWN
13	11	1069908	1305008	12000	12000	12000	36 month	12.69	402.54	B	B5	UCLA	10+ years	OWN
14	12	1064687	1298717	9000	9000	9000	36 month	13.49	305.38	C	C1	Va. Dept o	< 1 year	RENT
15	13	1069866	1304956	3000	3000	3000	36 month	9.91	96.68	B	B1	Target	3 years	RENT
16	14	1069057	1303503	10000	10000	10000	36 month	10.65	325.74	B	B2	SFMTA	3 years	RENT
17	15	1069759	1304871	1000	1000	1000	36 month	16.29	35.31	D	D1	Internal re	< 1 year	RENT
18	16	1065775	1299699	10000	10000	10000	36 month	15.27	347.98	C	C4	Chin's Res	4 years	RENT
19	17	1069971	1304884	3600	3600	3600	36 month	6.03	109.57	A	A1	Duracell	10+ years	MORTGAG
20	18	1062474	1294539	6000	6000	6000	36 month	11.71	198.46	B	B3	Connectio	1 year	MORTGAG
21	19	1069742	1304855	9200	9200	9200	36 month	6.03	280.01	A	A1	Network In	6 years	RENT
22	20	1069740	1284848	20250	20250	19142.16	60 month	15.27	484.63	C	C4	Archdioce	3 years	RENT
23	21	1039153	1269083	21000	21000	21000	36 month	12.42	701.73	B	B4	Osram Syl	10+ years	RENT
24	22	1069710	1304821	10000	10000	10000	36 month	11.71	330.76	B	B3	Value Air	10+ years	OWN

Exploratory Data Analysis

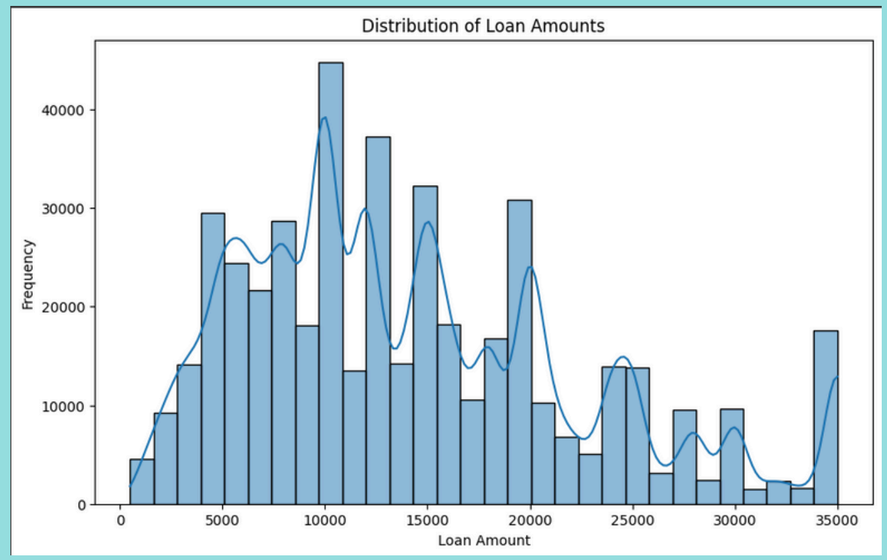
```
print(loan_df.shape)
```

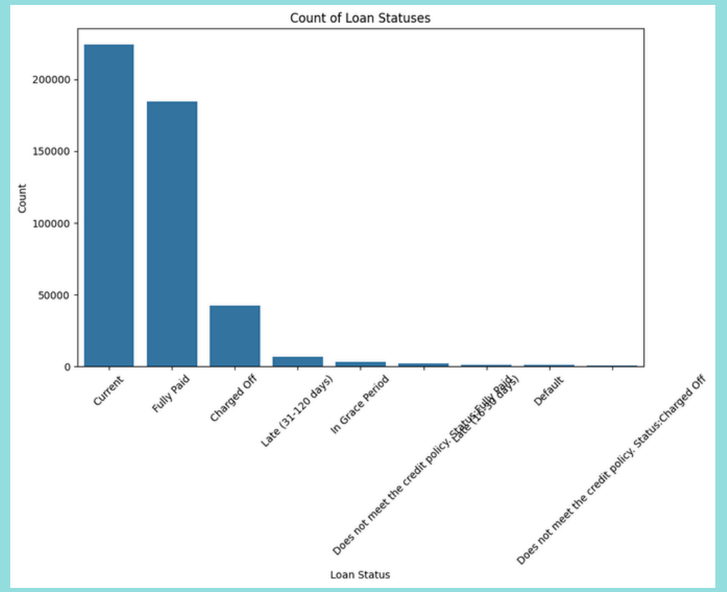
```
(466285, 75)
```

```
print(loan_df.describe())
```

```
          Unnamed: 0            id     member_id      loan_amnt  \
count  466285.000000  4.662850e+05  4.662850e+05  466285.000000
mean   233142.000000  1.307973e+07  1.459766e+07   14317.277577
std    134605.029472  1.089371e+07  1.168237e+07    8286.509164
min         0.000000  5.473400e+04  7.047300e+04     500.000000
25%    116571.000000  3.639987e+06  4.379705e+06    8000.000000
50%    233142.000000  1.010790e+07  1.194108e+07   12000.000000
75%    349713.000000  2.073121e+07  2.300154e+07   20000.000000
max    466284.000000  3.809811e+07  4.086083e+07   35000.000000

          funded_amnt  funded_amnt_inv       int_rate    installment  \
count  466285.000000    466285.000000  466285.000000  466285.000000
mean    14291.801044     14222.329888      13.829236     432.061201
std      8274.371300      8297.637788       4.357587     243.485550
min       500.000000         0.000000       5.420000      15.670000
25%      8000.000000      8000.000000      10.990000     256.690000
50%     12000.000000     12000.000000      13.660000     379.890000
75%     20000.000000     19950.000000      16.490000     566.580000
max     35000.000000     35000.000000      26.060000    1409.990000
```





Data Preparation

```
columns_to_drop = ['Unnamed: 0', 'id', 'member_id', 'emp_title', 'url', 'desc', 'title', 'zip_code', 'policy_code', 'addr_state', 'earliest_cr_line']

# Drop the columns
loan_df = loan_df.drop(columns=columns_to_drop)
```

```
loan_df.dropna(axis=1, how='all', inplace=True)


# Remove rows where loan_status is "Current"
loan_df = loan_df[loan_df['loan_status'] != 'Current']
```

```
loan_df['emp_length'] = loan_df['emp_length'].fillna(0)
loan_df['mths_since_last_delinq'] = loan_df['mths_since_last_delinq'].fillna(0)
loan_df['mths_since_last_record'] = loan_df['mths_since_last_record'].fillna(0)
loan_df['mths_since_last_major_derog'] = loan_df['mths_since_last_major_derog'].fillna(0)
```

```
loan_df['tot_coll_amt'] = loan_df['tot_coll_amt'].fillna(loan_df['tot_coll_amt'].mean())
loan_df['tot_cur_bal'] = loan_df['tot_cur_bal'].fillna(loan_df['tot_cur_bal'].mean())
loan_df['total_rev_hi_lim'] = loan_df['total_rev_hi_lim'].fillna(loan_df['total_rev_hi_lim'].mean())
```

```
X = loan_df.drop(columns=['loan_status', 'credit_risk'])  # Drop 'loan_status' if it's not needed
y = loan_df['credit_risk']


#Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (193460, 104) (193460,)
Test set: (48365, 104) (48365,)
```

Data Modelling

K-Nearest Neighbors

```
Results for k=3:
Training Data Performance:
              precision    recall  f1-score   support

           0       0.98      0.89      0.93     45766
           1       0.97      0.99      0.98    147694

    accuracy                           0.97    193460
   macro avg       0.97      0.94      0.96    193460
weighted avg       0.97      0.97      0.97    193460

Confusion Matrix on Training Data:
[[ 40528   5238]
 [   809 146885]]
AUC-ROC on Training Data: 0.995964274360998
Test Data Performance:
              precision    recall  f1-score   support

           0       0.95      0.81      0.87     11459
           1       0.94      0.99      0.96     36906

    accuracy                           0.94     48365
   macro avg       0.95      0.90      0.92     48365
weighted avg       0.94      0.94      0.94     48365

Confusion Matrix on Test Data:
[[ 9277  2182]
 [  520 36386]]
AUC-ROC on Test Data: 0.9447460852599121
Average accuracy of each hyperparameter variation (training data):
[1.         0.97747855 0.96874289]
Average accuracy of each hyperparameter variation (test data):
[0.93848858 0.9268686  0.94413315]
```

Logistic Regression

```
Best Parameters: {'C': 10, 'l1_ratio': 0.5, 'penalty': 'l2', 'solver': 'saga'}
Training Data Performance:
              precision    recall  f1-score   support

           0       0.99      0.78      0.87     45766
           1       0.94      1.00      0.97    147694

    accuracy                           0.95    193460
   macro avg       0.96      0.89      0.92    193460
weighted avg       0.95      0.95      0.94    193460

AUC-ROC on Training Data: 0.9798210445552472

Test Data Performance:
              precision    recall  f1-score   support

           0       0.99      0.78      0.88     11459
           1       0.94      1.00      0.97     36906

    accuracy                           0.95     48365
   macro avg       0.96      0.89      0.92     48365
weighted avg       0.95      0.95      0.95     48365

AUC-ROC on Test Data: 0.9801779972523151
```

Evaluation

K-Nearest Neighbors

Training Accuracy: 96%
Testing Accuracy: 94%
Slightly Overfitting

Logistic Regression

Training Accuracy: 97%
Testing Accuracy: 98%
Goodfitting

Thank You