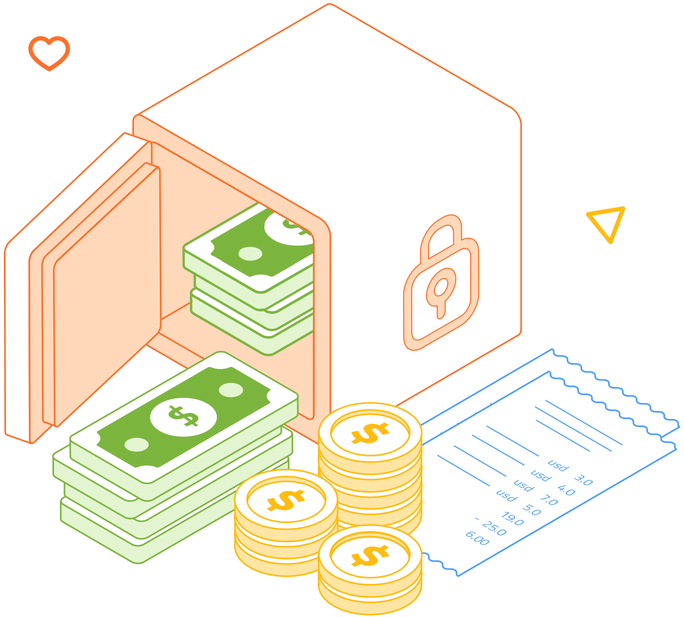


CREDIT DEFAULT PREDICTION

By: Ahmad Zaki Irfan

Table of contents

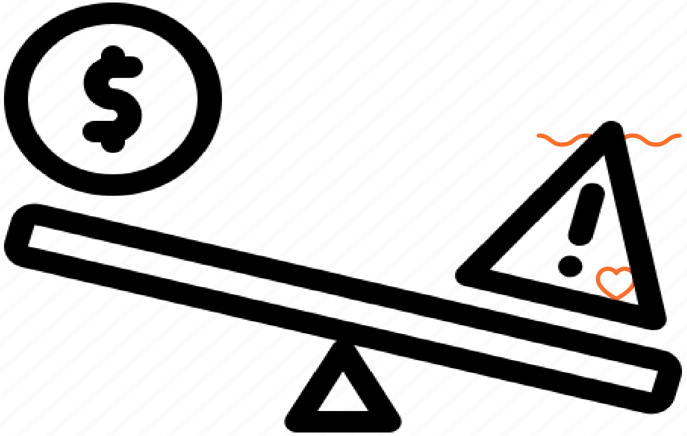
Background & Problem Statement

01

1. Background & Problem Statement

Credit default is a risk that must be minimized in the Banking institution because the more credit default are, the greater the loss reserves that must be formed by the Bank to be able to cover these credit default, so that it will have an impact on the Bank's profit and loss.

Therefore, the Bank needs to mitigate the risk of credit default. By predicting it, the mitigation will be more effective and efficient.



1. Background & Problem Statement

The purpose of this project is to create a credit default prediction model so that the Bank can do early risk mitigation.

Methodology

02

2. Methodology

Data Understanding → Data Cleansing → EDA → Data Preprocessing → Modeling (Baseline + Improvement) → Model Interpretation → Business Recommendation

Data Understanding & Data Cleansing

03

3. Data Understanding & Data Cleansing

Data source : *https://www.kaggle.com/datasets/laotse/credit-risk-dataset*
(32,581 rows, 12 column)

No	Feature Name	Description	Dtype
1	person_age	Age	Numerical
2	person_income	Annual Income	Numerical
3	person*home*ownership	Home ownership	Categorical
4	person*emp*length	Employment length (in years)	Numerical
5	loan_intent	Loan intent	Categorical
6	loan_grade	Loan grade	Categorical
7	loan_amnt	Loan amount	Numerical
8	Loan*int*rate	Interest rate	Numerical
9	loan_status	Loan status (0 is non default 1 is default)	Numerical
10	loan*percent*income	Percent income	Numerical
11	cb*person*default*on*file	Historical default	Categorical
12	cb*preson*cred*hist*length	Credit history length	Numerical

3. Data Understanding & Data Cleansing

Missing Value

```python
def missing_value_check (df) :
    null_number = (df.isnull().sum()/len(df))*100
    return null_number.sort_values(ascending = False)
missing_value_check(df)
```

```
loan_int_rate                 9.563856
person_emp_length             2.747000
person_age                    0.000000
person_income                 0.000000
person_home_ownership         0.000000
loan_intent                   0.000000
loan_grade                    0.000000
loan_amnt                     0.000000
loan_status                   0.000000
loan_percent_income           0.000000
cb_person_default_on_file     0.000000
cb_person_cred_hist_length    0.000000
dtype: float64
```

1. The missing value in the person_emp_length feature will be dropped because the value is not significant to the dataset
2. The missing value in the loan_int_rate feature will be imputed using mean

3. Data Understanding & Data Cleansing

```
[44] df_n.duplicated().sum()
157
```

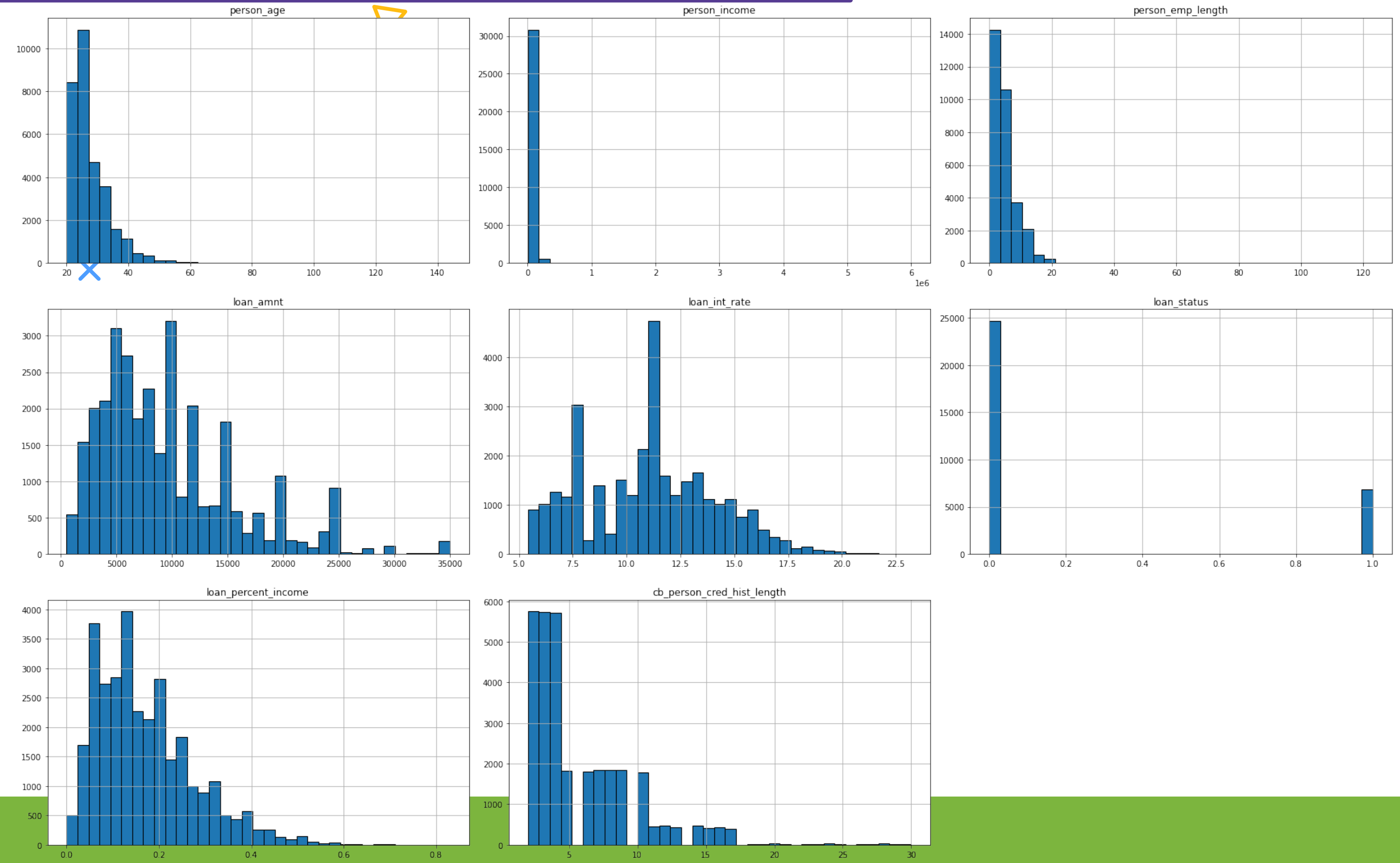
Duplicated Value

> Drop 157 duplicated values, so new dataset without missing value and duplicated value becomes ***31,482 rows and 12 columns*** or ***97%*** from original dataset.

Exploratory Data Analysis

04

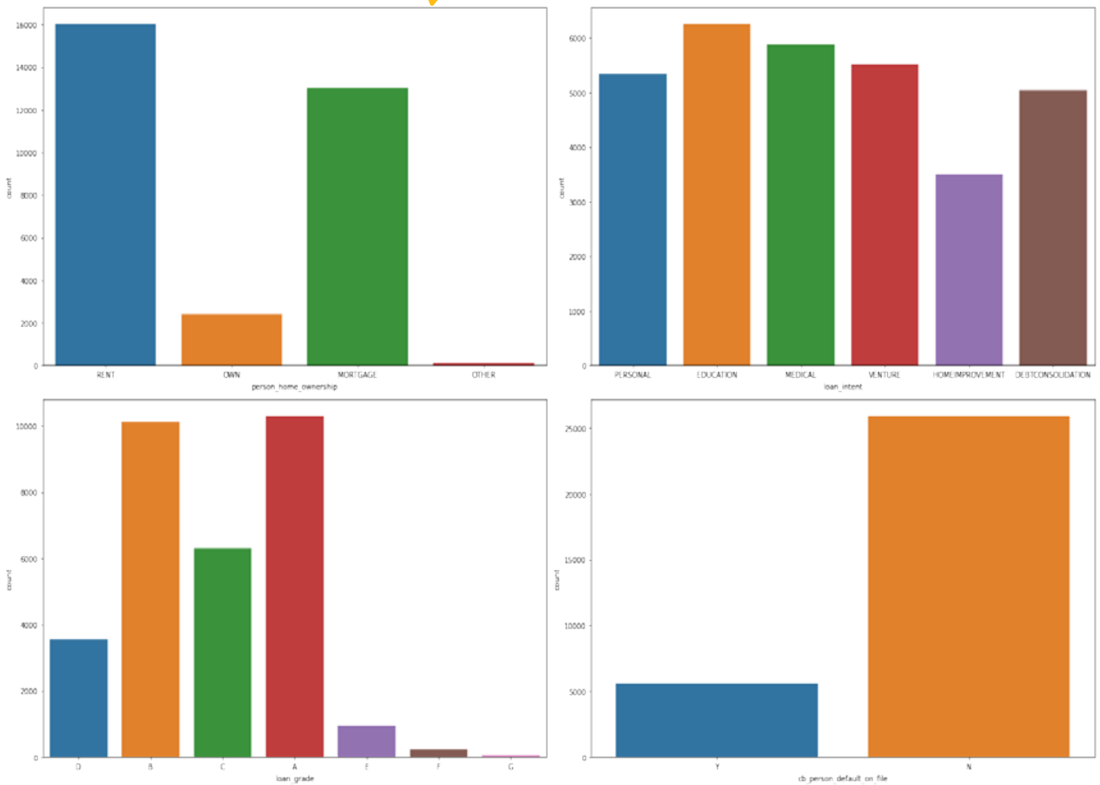
4. Exploratory Data Analysis



4. Exploratory Data Analysis

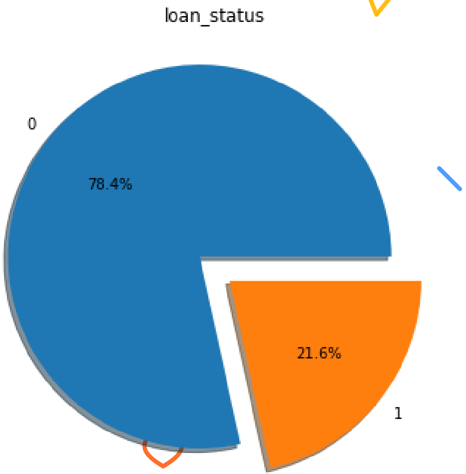
1. There are discrepancies in the maximum value in the person_age, person_income and person_emp_length features because the maximum value of each feature is too far from the average value
2. The average customer income is 66,074 per year
3. The average customer works it has been for 5 years
4. The average number of customer loans is 9,589
5. The average interest rate on customer loans is 11% per year
6. The average number of customer loans is 17% of customer income per year
7. The average customer have credit history that has been running for 6 months

4. Exploratory Data Analysis

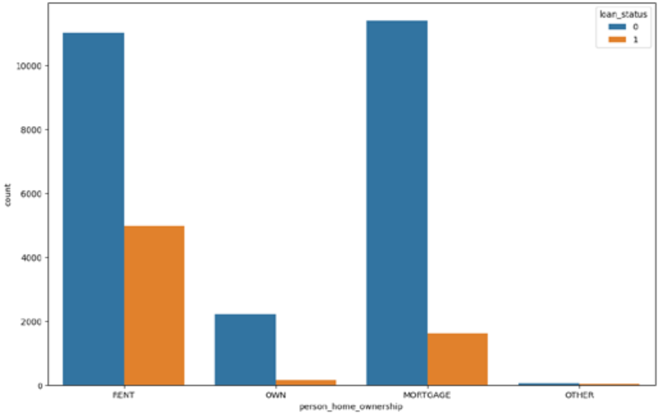


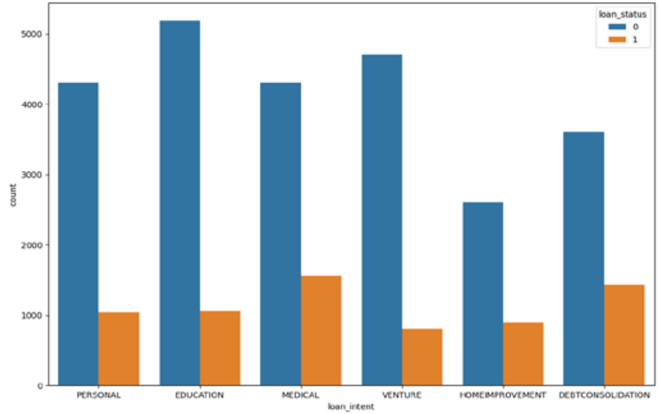
1. Most of the customer's residences are still rented
2. The highest purpose of using credit is for Education
3. Highest loan grade for customers is grade A
4. Most of the customers have no history of credit default before

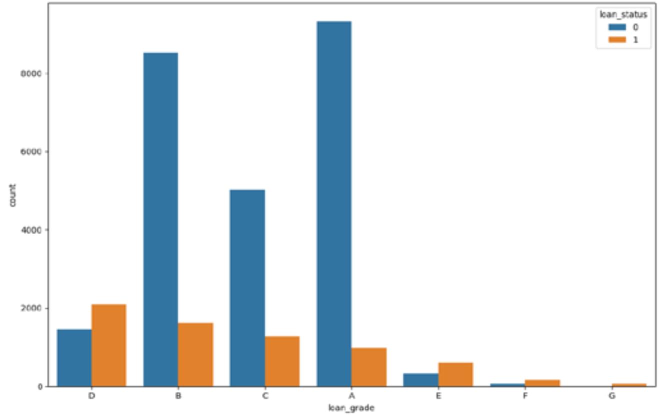
4. Exploratory Data Analysis

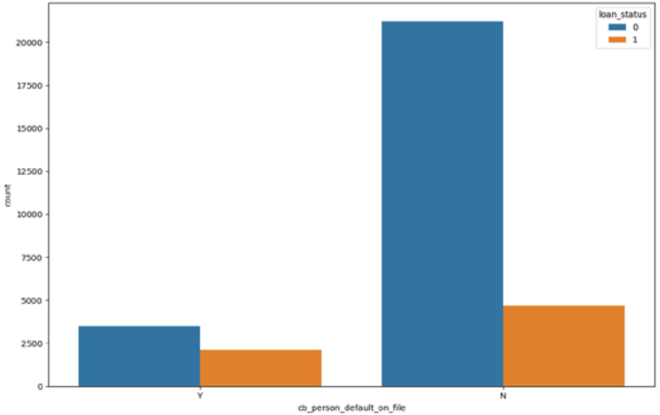


21.8% or 7089 credit is default









4. Exploratory Data Analysis

Credit Default Customers Profile

age
average 27 years old

income
average 50,000/year

employment length
average 4 years

loan amount
average 11,000

loan rate
Rata-rata bunga kredit 13% / tahun

loan/income
average 24%

credit history
the average previous credit has been running for 5 months

home ownership
73% rent

loan intent
23% for health

loan grade
30% grade “D”

default history
69% with no default history

Data Preprocessing & Modeling

05

Data Preprocessing

01 Abnormal Data Handling

1. The maximum value in the person_age feature (age > 100 years) and person_emp_length feature (work experience > 100 years) is too far from the average value.
2. Age > 64 years or exceeding the maximum productive age assumption at 64 years of age.
3. Work experience > 49 years
4. Age starting work <15 year or > 64 years

02 Encoding Categorical Feature

1. Label Encoding
2. One Hot Encoding
3. Rank Encoding

03 Stratified Sampling

1. Data Train : 70%
2. Data Test : 30%

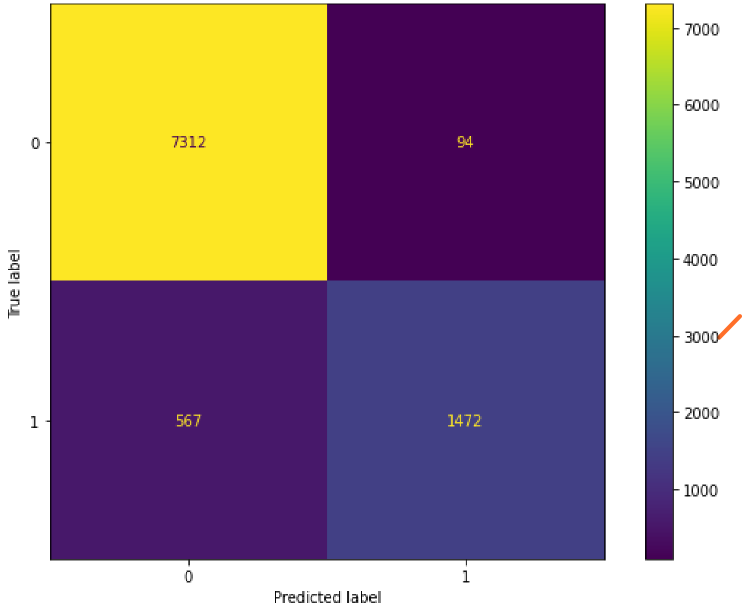
04 Balancing Data

Oversampling SMOTE

Modeling – Baseline Model

	f1_score_baseline_model
Random Forest Classification	0.816644
Decision Tree Classification	0.728862
Logistic Regression	0.521724

Random Forest has the highest f1 score compared to other models, but it needs to be improved by tuning the parameters to get maximum results



	precision	recall	f1-score	support
0	0.93	0.99	0.96	7406
1	0.94	0.72	0.82	2039
accuracy			0.93	9445
macro avg	0.93	0.85	0.89	9445
weighted avg	0.93	0.93	0.93	9445

Modeling – Improvement Model

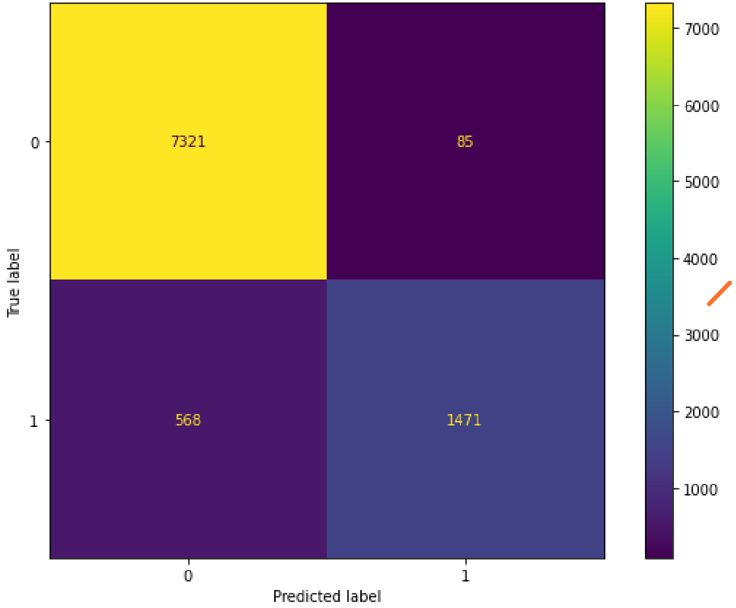
	f1_score_baseline_model	f1_score_improvement_model
Random Forest Classification	0.816644	0.818359
Decision Tree Classification	0.728862	0.739013
Logistic Regression	0.521724	0.623970

Chosen parameters model

```
rf_clf_gridcv.best_params_
```

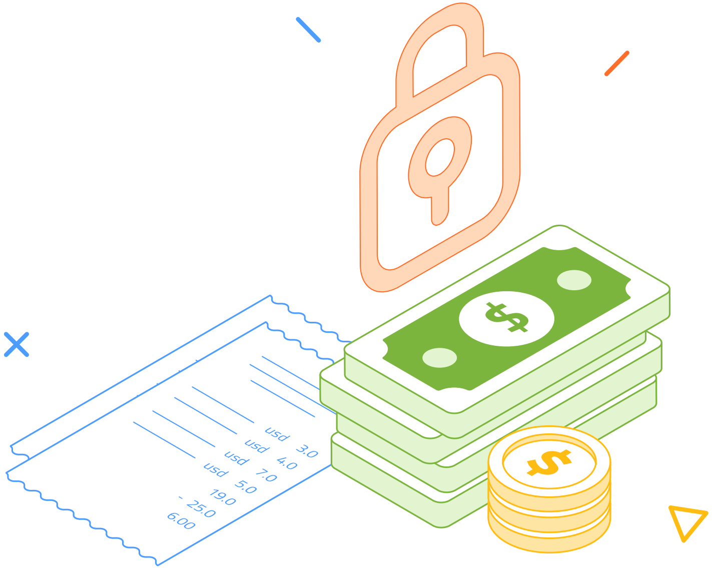
```
{'bootstrap': False,
 'criterion': 'entropy',
 'max_depth': None,
 'min_samples_leaf': 1,
 'min_samples_split': 4,
 'n_estimators': 100}
```

The Random Forest Improvement model has the highest f1 score compared to the baseline model and other improvement models, so the random forest improvement model is used to predict credit default.



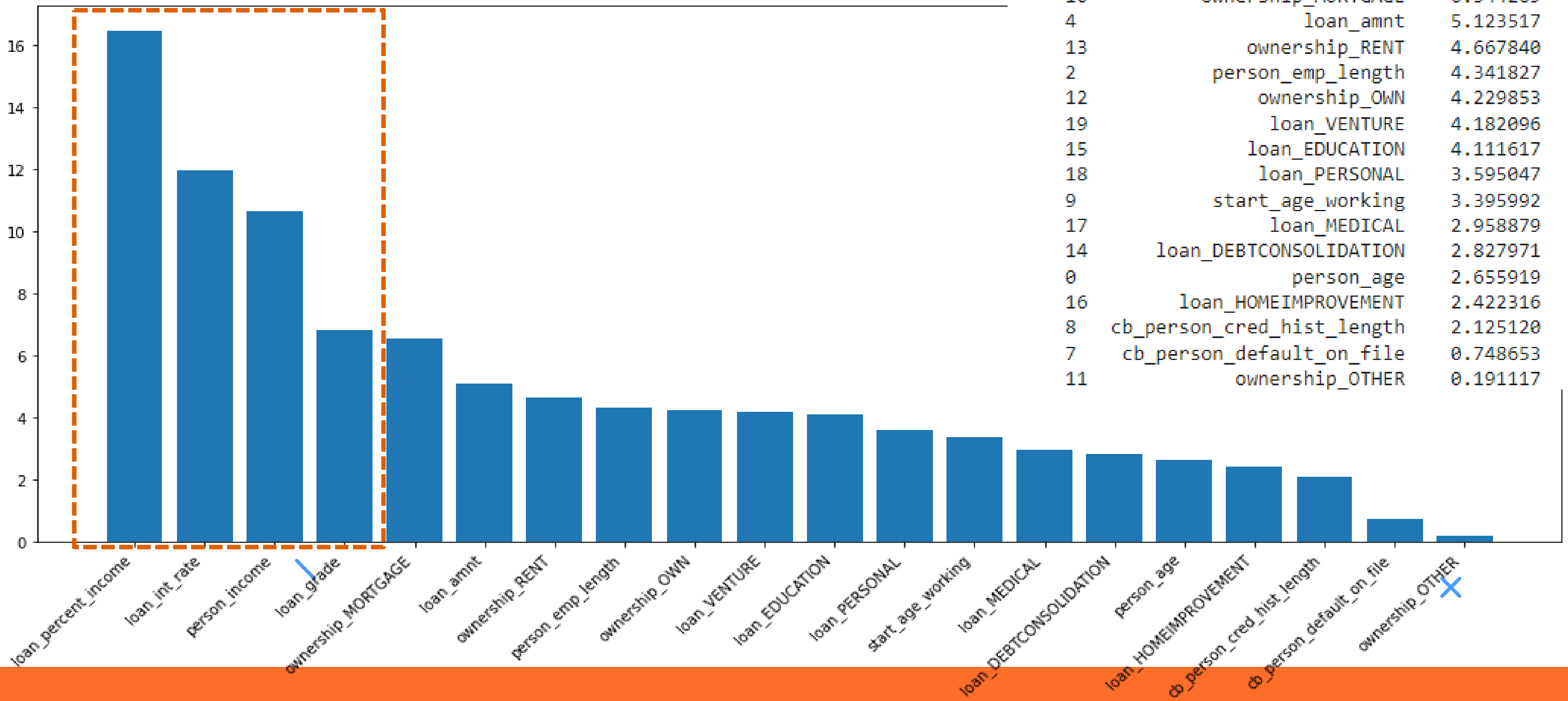
	precision	recall	f1-score	support
0	0.93	0.99	0.96	7406
1	0.95	0.72	0.82	2039
accuracy			0.93	9445
macro avg	0.94	0.85	0.89	9445
weighted avg	0.93	0.93	0.93	9445

Model Interpretation & Business Recommendation



06

Model Interpretation - Feature Importance (Top 4)



	Variable	Importance
6	loan_percent_income	16.468261
5	loan_int_rate	11.954882
1	person_income	10.635530
3	loan_grade	6.819295
10	ownership_MORTGAGE	6.544269
4	loan_amnt	5.123517
13	ownership_RENT	4.667840
2	person_emp_length	4.341827
12	ownership_OWN	4.229853
19	loan_VENTURE	4.182096
15	loan_EDUCATION	4.111617
18	loan_PERSONAL	3.595047
9	start_age_working	3.395992
17	loan_MEDICAL	2.958879
14	loan_DEBTCONSOLIDATION	2.827971
0	person_age	2.655919
16	loan_HOMEIMPROVEMENT	2.422316
8	cb_person_cred_hist_length	2.125120
7	cb_person_default_on_file	0.748653
11	ownership_OTHER	0.191117

Model Interpretation - Feature Contribution

```
from treeinterpreter import treeinterpreter as ti
import waterfall_chart

def create_contrbutions_df(row):
    row_value = X_test_smote.values[[row]]
    prediction, bias, contributions = ti.predict(rf_clf_gridcv.best_estimator_, row_value)
    idxs = np.argsort(contributions[0][:][:,1])
    contrib_df = pd.DataFrame([o for o in zip(X_test_smote.columns[idxs], X_test_smote.iloc[row][idxs], contributions[0][:][idxs,1])])
    pred = contrib_df[2].sum()+bias[0][0]
    print (contrib_df)
    print ("bias :", bias[0][0])
    print ("contributions :", contrib_df[2].sum())
    print ("calculated prediction :", pred)
    print("final model prediction :",rf_clf_gridcv.best_estimator_.predict(X_test_smote.values[[row]])[0])
    plt.rcParams.update({'figure.figsize':(7.5,5), 'figure.dpi':100})
    my_plot=waterfall_chart.plot(contrib_df[0],contrib_df[2],sorted_value= True, rotation_value=90, threshold=0.1,formatting='{:,.3f}')
```

This function shows the contribution value of each feature to the decision model for predicting credit default or not by using waterfall chart

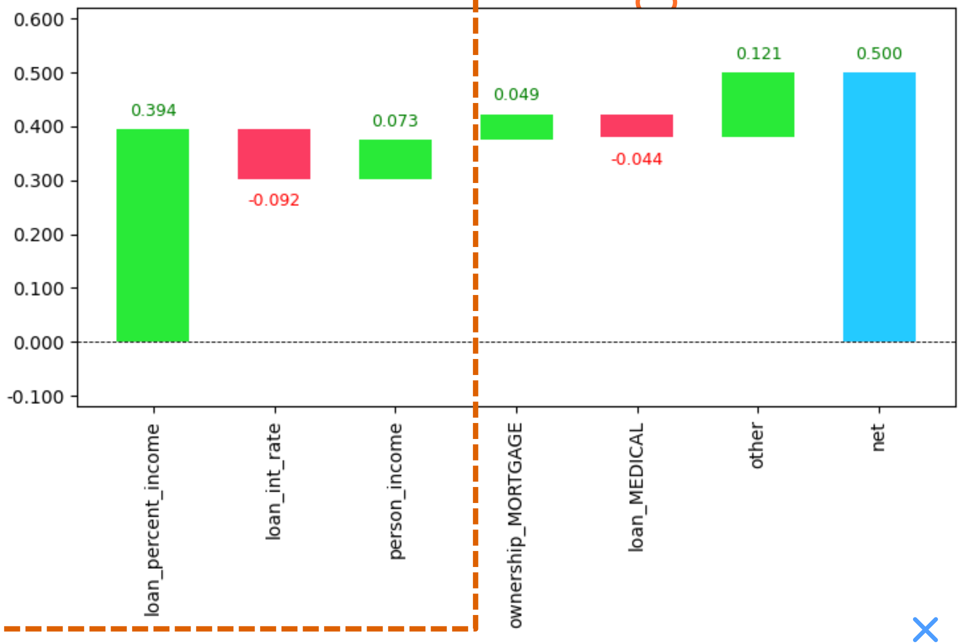
Model Interpretation - Feature Contribution (1)

Customer no. 33 in testing data

```
create_contrbutions_df(33)
```

	0	1	2
0	loan_int_rate	6.54	-0.092493
1	loan_MEDICAL	1.00	-0.043816
2	loan_grade	1.00	-0.021724
3	person_age	25.00	-0.000200
4	ownership_OTHER	0.00	0.000000
5	cb_person_default_on_file	0.00	0.000024
6	loan_HOMEIMPROVEMENT	0.00	0.001756
7	loan_DEBTCONSOLIDATION	0.00	0.001902
8	loan_PERSONAL	0.00	0.003993
9	cb_person_cred_hist_length	2.00	0.004670
10	person_emp_length	1.00	0.006570
11	start_age_working	24.00	0.007645
12	loan_EDUCATION	0.00	0.011709
13	loan_VENTURE	0.00	0.011937
14	ownership_OWN	0.00	0.021719
15	loan_amnt	9250.00	0.032188
16	ownership_RENT	1.00	0.038748
17	ownership_MORTGAGE	0.00	0.048788
18	person_income	25716.00	0.072503
19	loan_percent_income	0.36	0.394083

```
bias : 0.5
contributions : 0.4999999999999999
calculated prediction : 0.9999999999999999
final model prediction : 1
```



Top 4 feature has more contribution

Model Interpretation - Feature Contribution (2)

Customer no. 16 in testing data

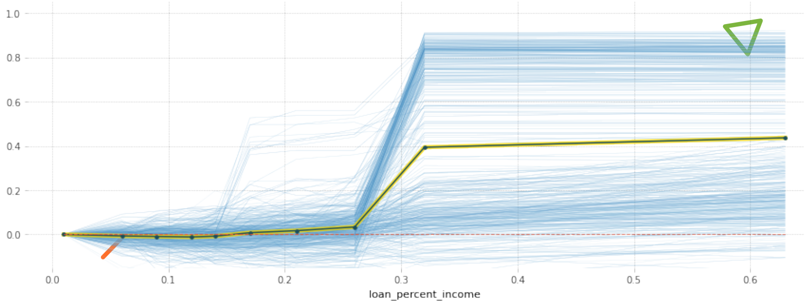
```
create_contrbutions_df(16)
```

	0	1	2
0	loan_percent_income	0.06	-0.076548
1	loan_HOMEIMPROVEMENT	1.00	-0.058124
2	cb_person_cred_hist_length	16.00	-0.021929
3	cb_person_default_on_file	1.00	-0.014978
4	person_age	47.00	-0.001481
5	ownership_OTHER	0.00	0.000032
6	start_age_working	47.00	0.001316
7	loan_DEBTCONSOLIDATION	0.00	0.002245
8	loan_MEDICAL	0.00	0.006889
9	loan_amnt	1000.00	0.007947
10	ownership_RENT	1.00	0.015677
11	loan_PERSONAL	0.00	0.017172
12	loan_VENTURE	0.00	0.020795
13	ownership_OWN	0.00	0.027586
14	loan_EDUCATION	0.00	0.028112
15	person_emp_length	0.00	0.045652
16	ownership_MORTGAGE	0.00	0.053505
17	person_income	18000.00	0.106291
18	loan_int_rate	14.84	0.136662
19	loan_grade	4.00	0.161512

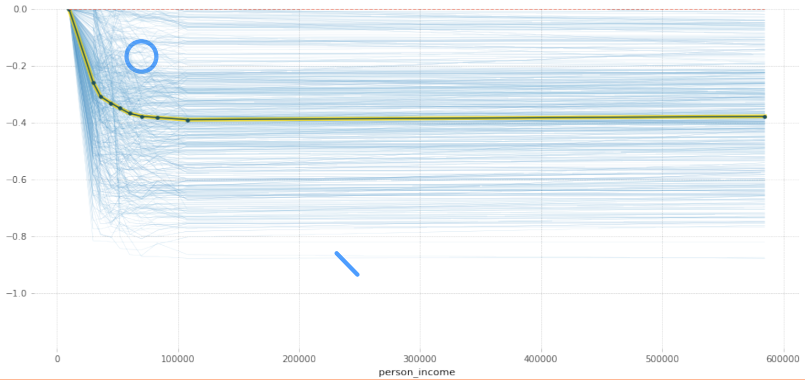
```
bias : 0.5
contributions : 0.4583333333333326
calculated prediction : 0.9583333333333333
final model prediction : 1
```

Top 4 feature has more contribution

Model Interpretation - Partial Dependence (Top 4 Feature Importance)



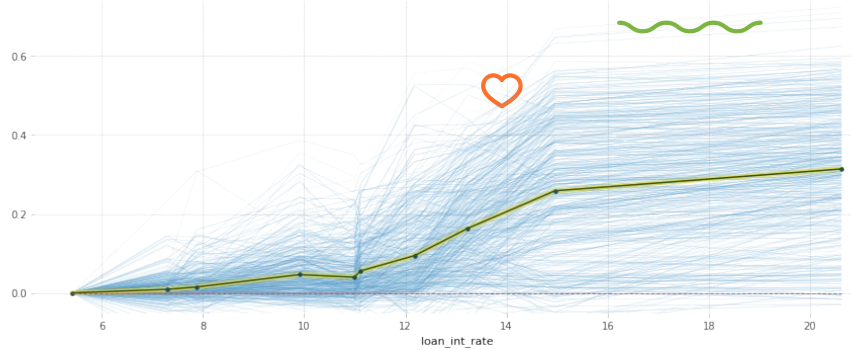
customers will tend to default when credit/income is in the position of 24% and above



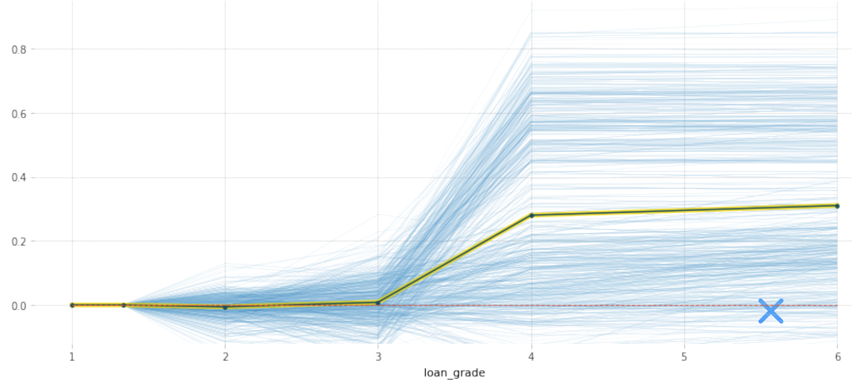
the greater the customer's income, they will not tend to default

Model Interpretation - Partial Dependence (Top 4 Feature Importance)

customers will tend to default when interest is in the position of 11%/year and above



customers will tend to default when loan grades D,E,F,G



Business Recommendation

1. Bank can focus on 3 important features when analyzing credit, these features are:
   a. **loan/income**: Bank must be more careful with customers who have a credit/income value >= 24%
   b. **bunga**: Bank must be more careful with customers who have interest rates >= 11%
   c. **loangrade**: Bank must be more careful with customers who have loan grades D, E, F and G

2. The Recall score of model is 72%, meaning that out of 10 customers who are default, there are 3 customers that model failed to predict. In other words, the effectiveness of the Bank's loss reserves which is formed to be able to cover credit default reaches 72% so that financial allocation arrangements become more measurable.

3. Bank can carry out risk mitigation on loss reserves formed by sharing risks with 3rd parties, namely Insurance Institutions or Credit Guarantee Institutions for the criteria of customers who tend to default.

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**