CREDIT SCORING

by

Aria Dipa Itna Erlangga

AGENDA

Business understanding

Primary goals

Data preparation

EDA

Data preprocessing

Modeling & Model evaluation

BUSINESS UNDERSTANDING

A credit score is based on credit history: the number of open accounts, total levels of debt, repayment history, and other factors. Lenders use credit scores to evaluate the probability that an individual will repay loans in a timely manner

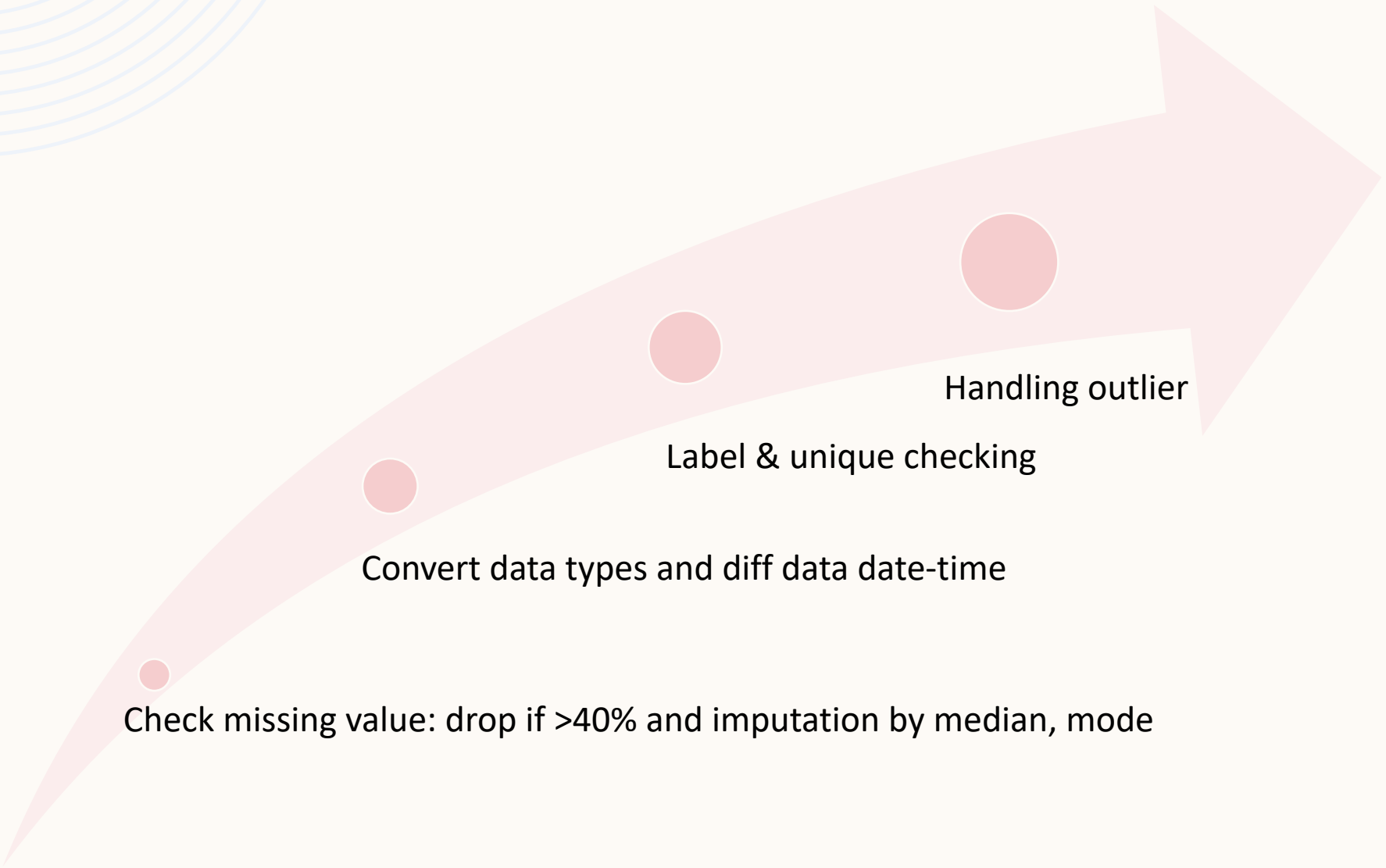
PRIMARY GOALS

- minimize the possibility of default
- determine the potential target market
- determine attributes that influence credit score

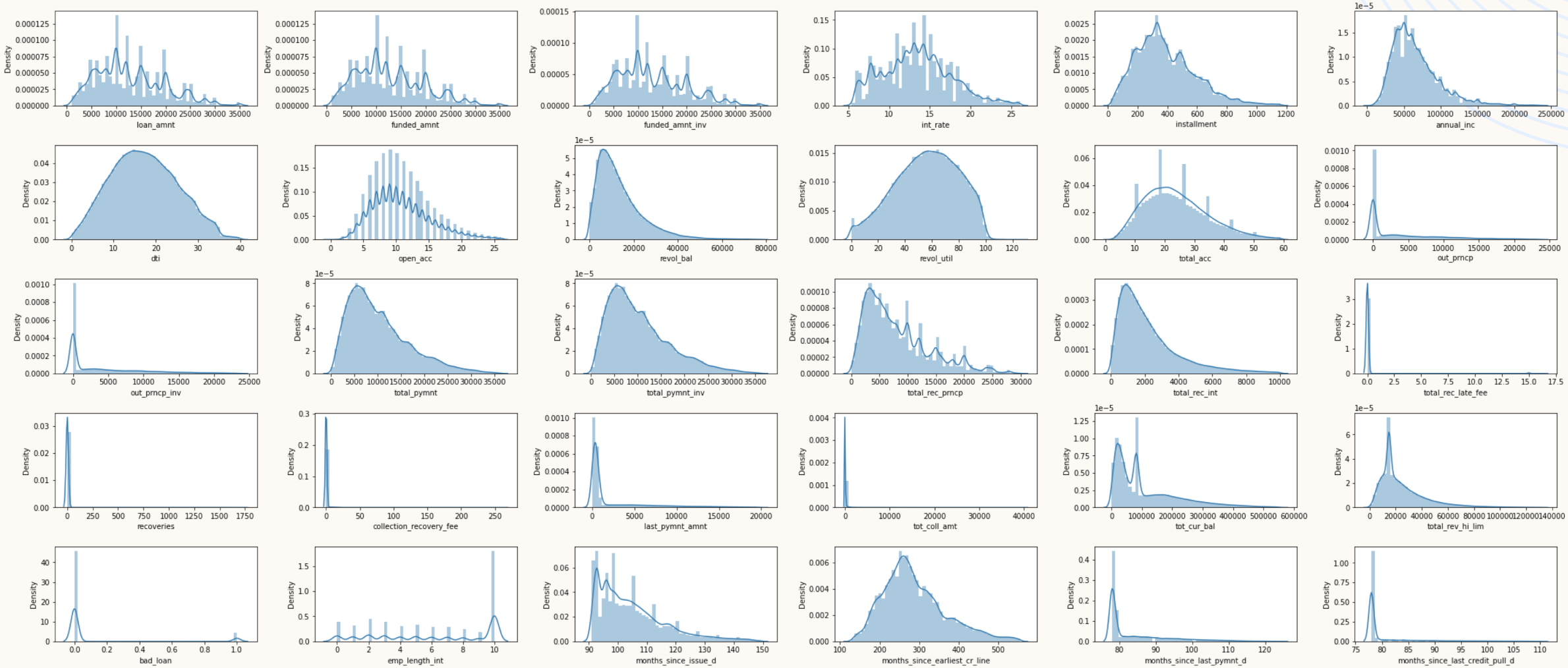
DATA UNDERSTANDING

- 466285 records, 72 Columns
- Imbalanced Target Labels
- dtypes: float64(46), int64(4), object(22)

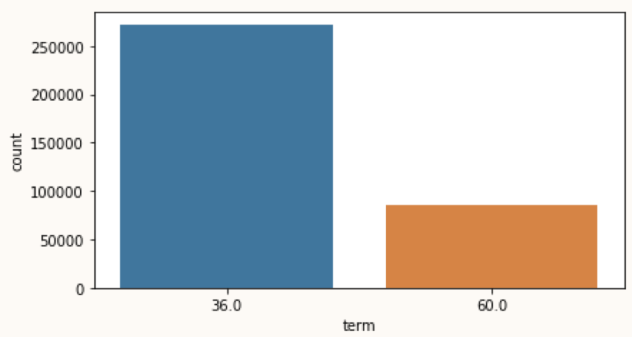
DATA PREPARATION

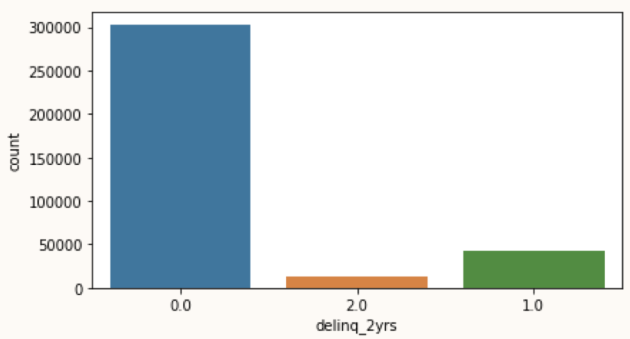


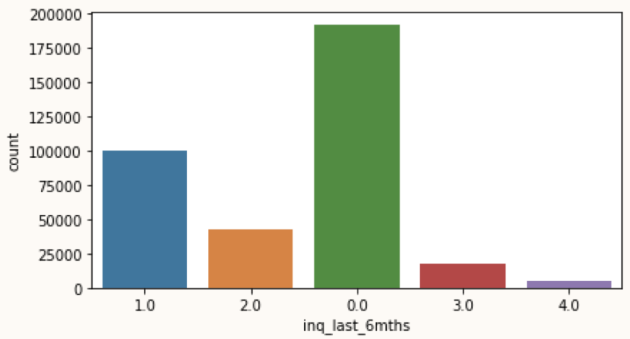
EDA

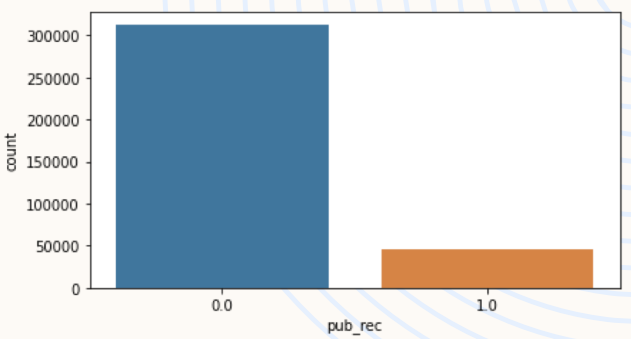


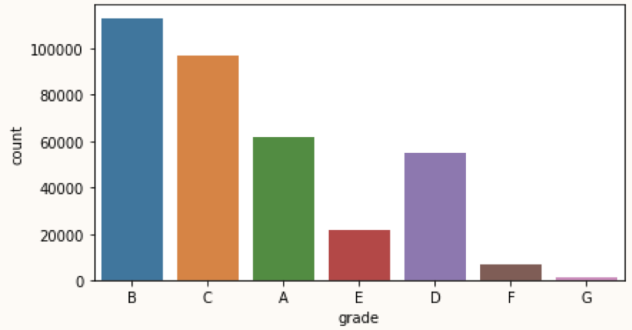
- distplot for all numeric variables and it can be seen that some variables are skew and multimodal for variables with high
- variance and high bias will be converted into categorical form | Q1=Q2=Q3=0

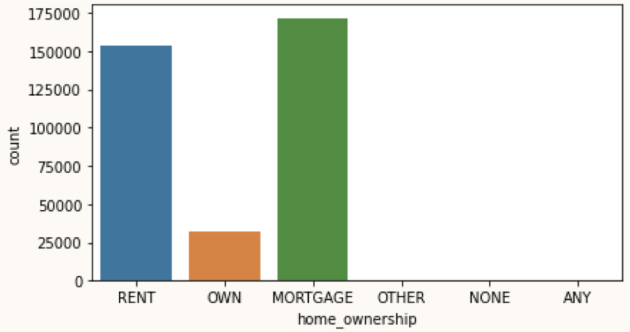


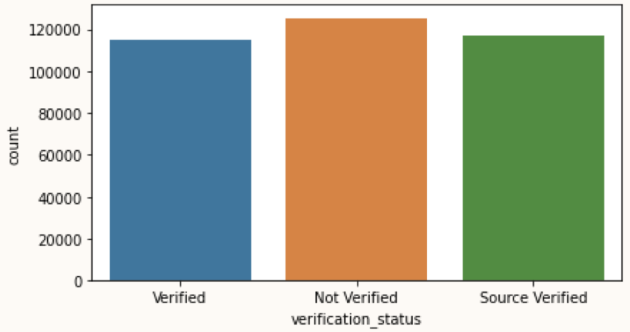


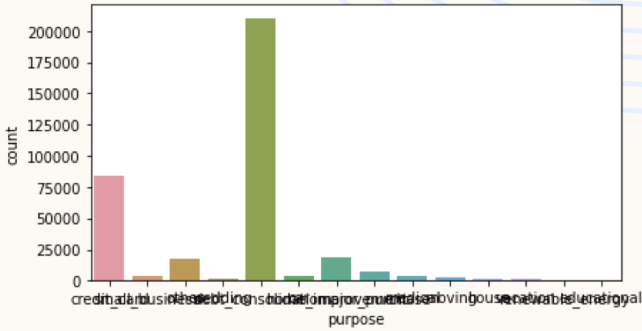


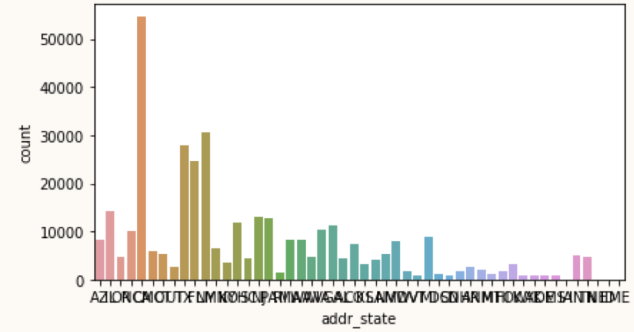


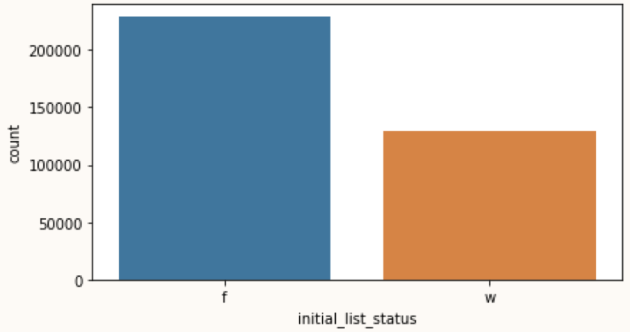


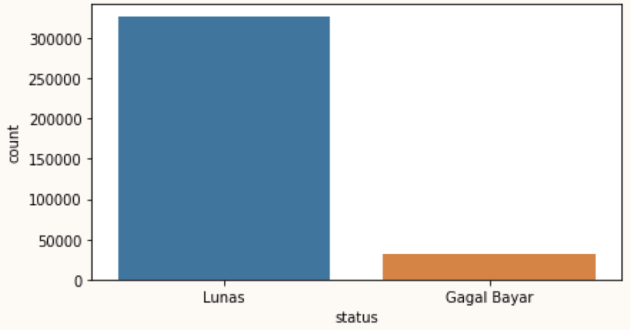








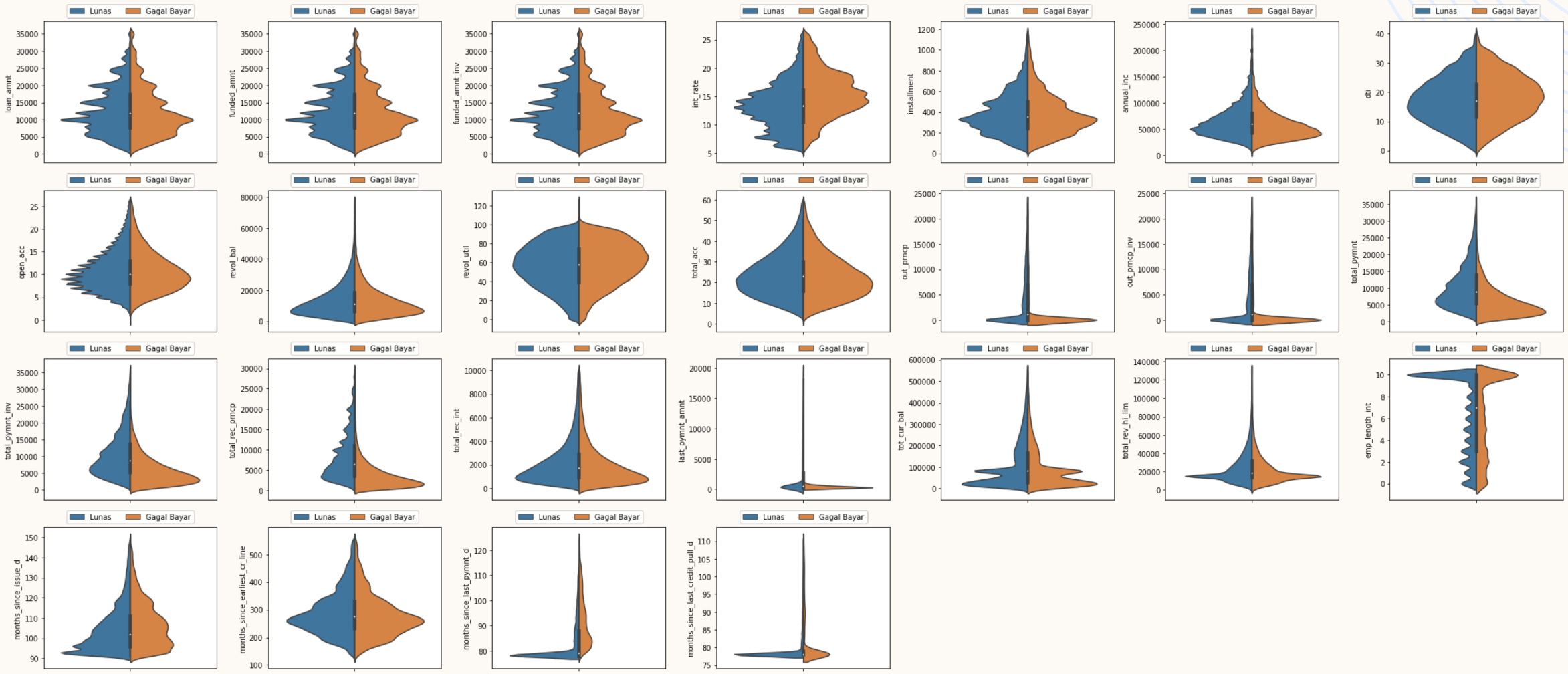




>>> It can be seen that the predictor variables with labels are not imbalanced and this will affect the accuracy of model, so a few labels are regrouped into several available labels or create other labels

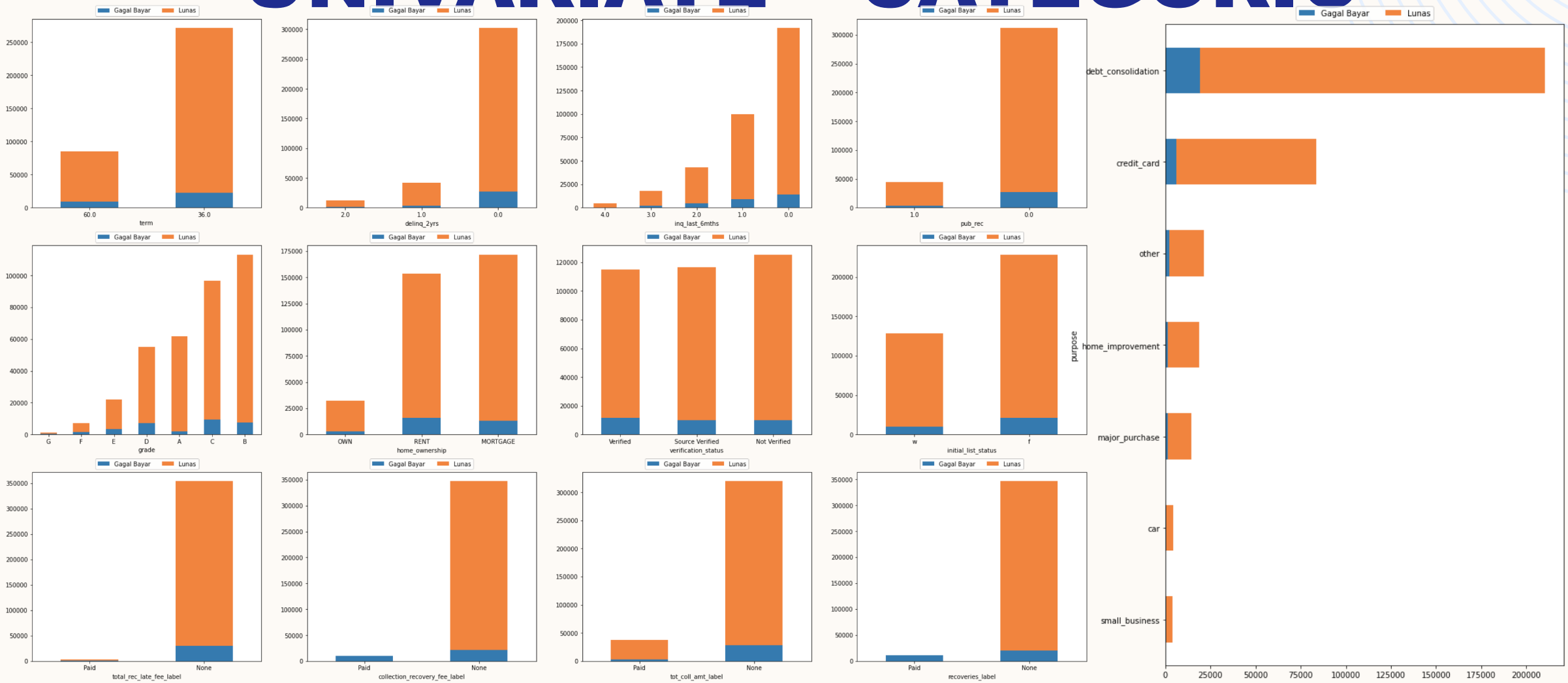
```
[ ]  df['home_ownership'].replace({'NONE':'RENT', 'ANY':'RENT', 'OTHER':'RENT'},inplace=True)

     df['purpose'].replace({'educational':'major_purchase',
                            'house':'major_purchase',
                            'medical':'major_purchase',
                            'moving':'major_purchase',
                            'vacation':'other',
                            'wedding':'other',
                            'renewable_energy':'home_improvement'},inplace=True)

     df['addr_state'].replace({'IA':'OTHER', 'ID':'OTHER', 'NE':'OTHER', 'ME':'OTHER'},inplace=True)
```

UNIVARIATE – NUMERIC



- dist plot for all numeric variables with paid and non-paid labels. The pattern of two labels so approaching
- there are several variables with high variance and high bias are maintained for log transformation

UNIVARIATE – CATEGORIC



- count plot for all categorical variables. There are some labels of imbalance variables and will be maintained
- for 4 plots in bottom are the conversion results from numeric variable

BIVARIATE - MULTIVARIATE

ANOVA

Find that for all numerical predictor variables of significance < 0.05, so H0 is rejected, means that there is an average difference between the tested target labels. Conclusion = there is an effect of all numerical predictor variables on credit status.

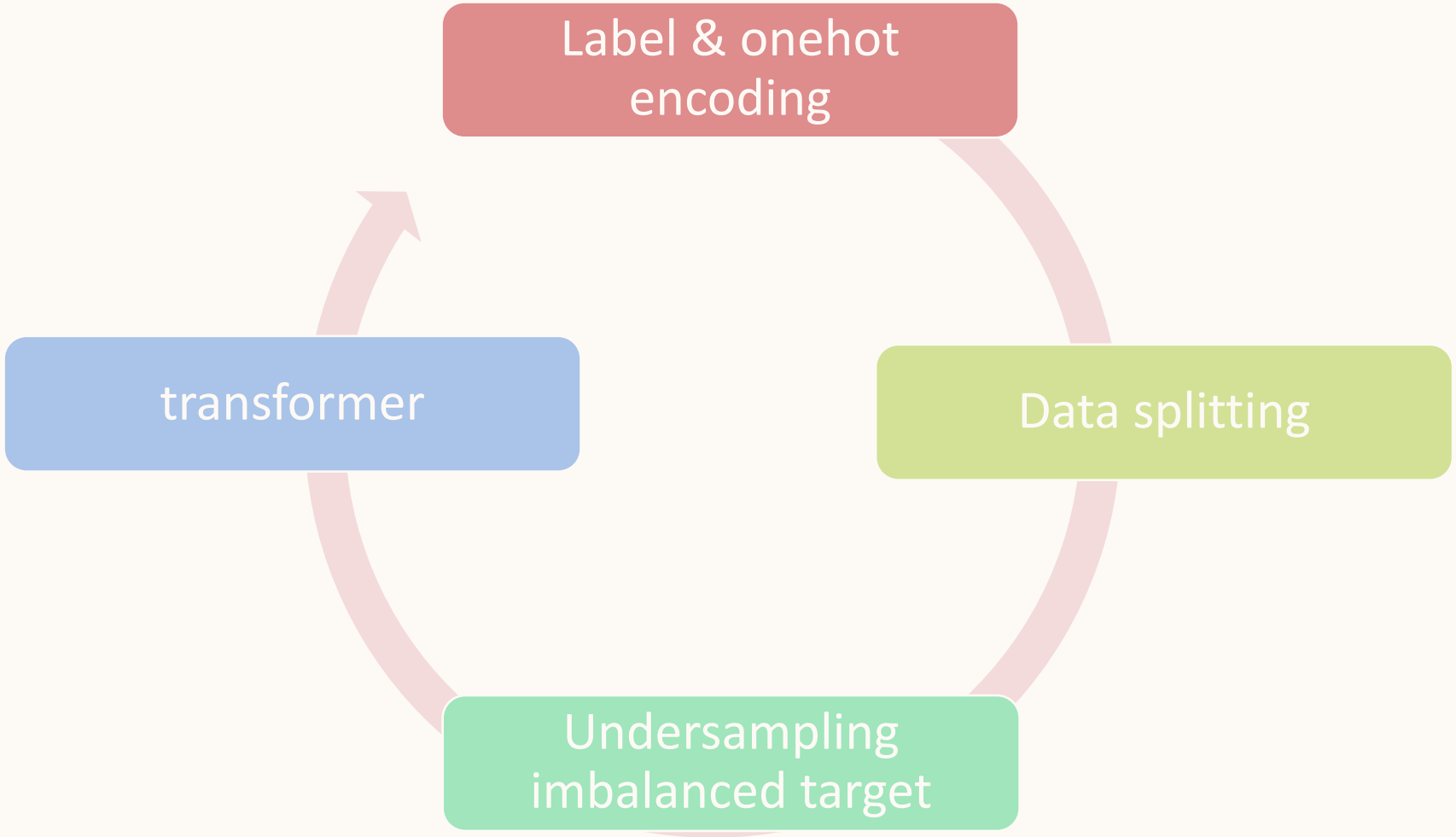
CHI-SQUARED

Find that almost all categorical predictor variables [except: pub_rec] sign < 0.05, so H0 is rejected, means that there is an average difference between the tested target labels. Conclusion = there is a categorical predictor variable has no effect on credit status and the variable will be deleted.

PEARSON CORR

Find that for correlation all numerical predictor variables and there are several variables that have a correlation < 0.7 and the variables will be deleted

DATA PREPROCESSING



MODELING & COMPARISON

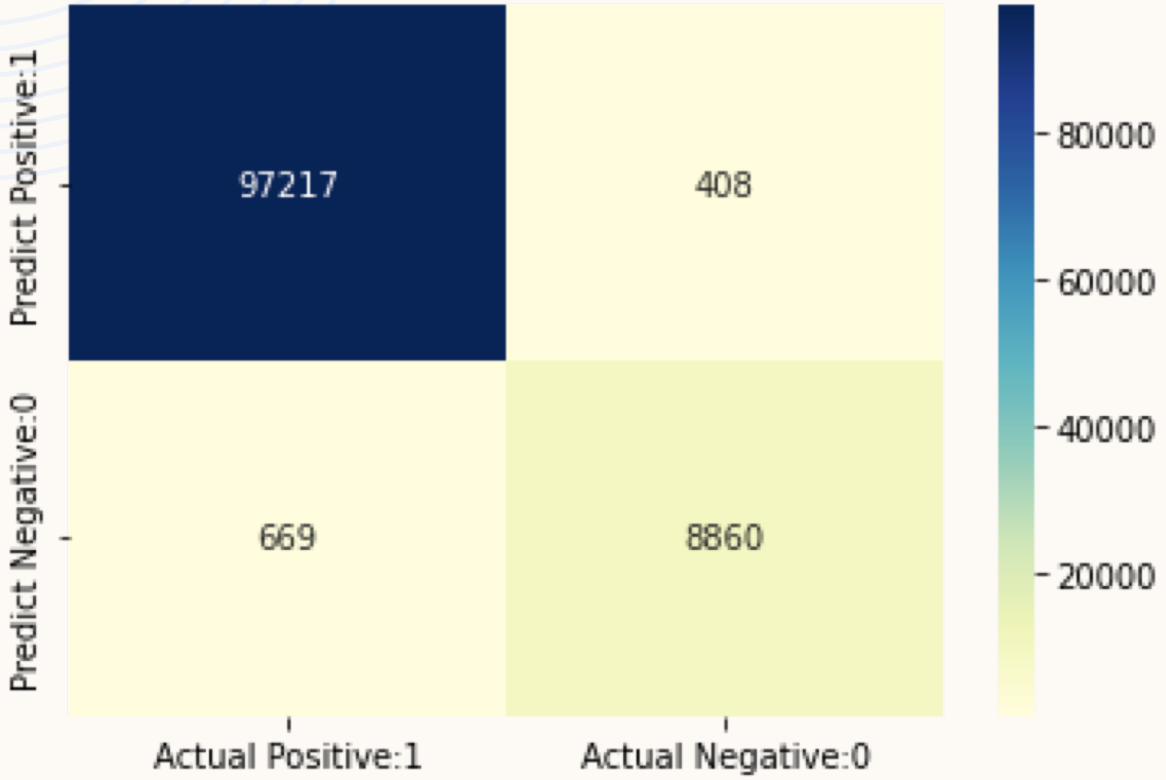
	Model	Akurasi Train	Akurasi Test
0	RandomForest	1.000000	0.988820
2	Xtree	1.000000	0.977248
3	DecisionTree	1.000000	0.984350
4	GradientBoost	0.972204	0.975139
1	LogisticRegression	0.888233	0.926862

It shown that RandomForest is the best algorithm from comparison on the left

```
Akurasi Train 1.0
Akurasi Test 0.9899490452992888
```

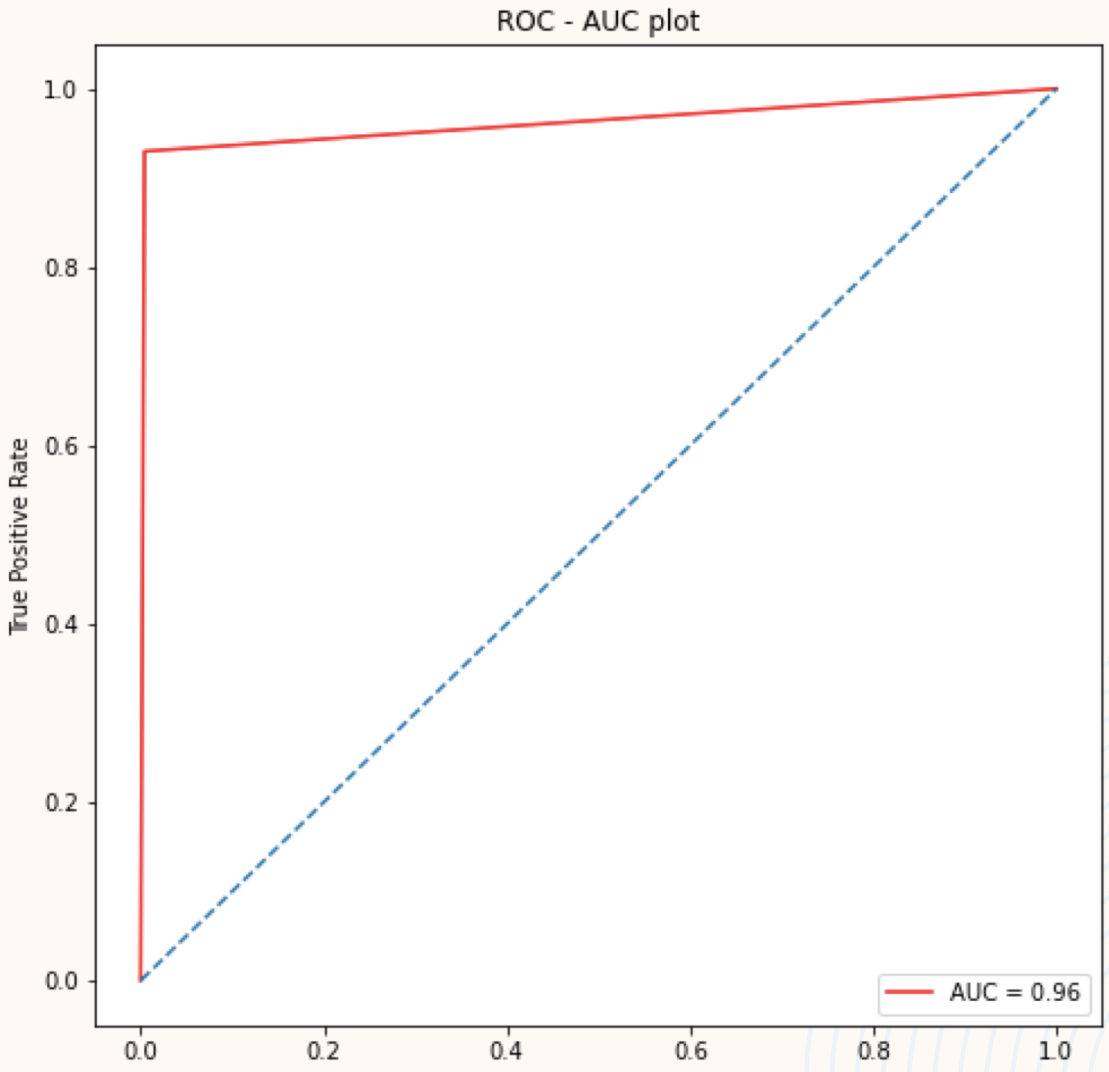
Model is not prone to underfit or overfit

MODEL EVALUATION



Akurasi Klasifikasi : 0.9899
Kesalahan Klasifikasi : 0.0101
Presisi : 0.9958
Sensitifitas : 0.9932

ROC – AUC >0.9 is considered outstanding classifier



THANK YOU

Aria Dipa Itna Erlangga
anggaitna@gmail.com
www.github.com/ariadipa