Credit Risk Evaluation using Machine Learning

By *Surya* Vaddhiparthy

Under the guidance of
Professor Dr. Guan Yue Hong

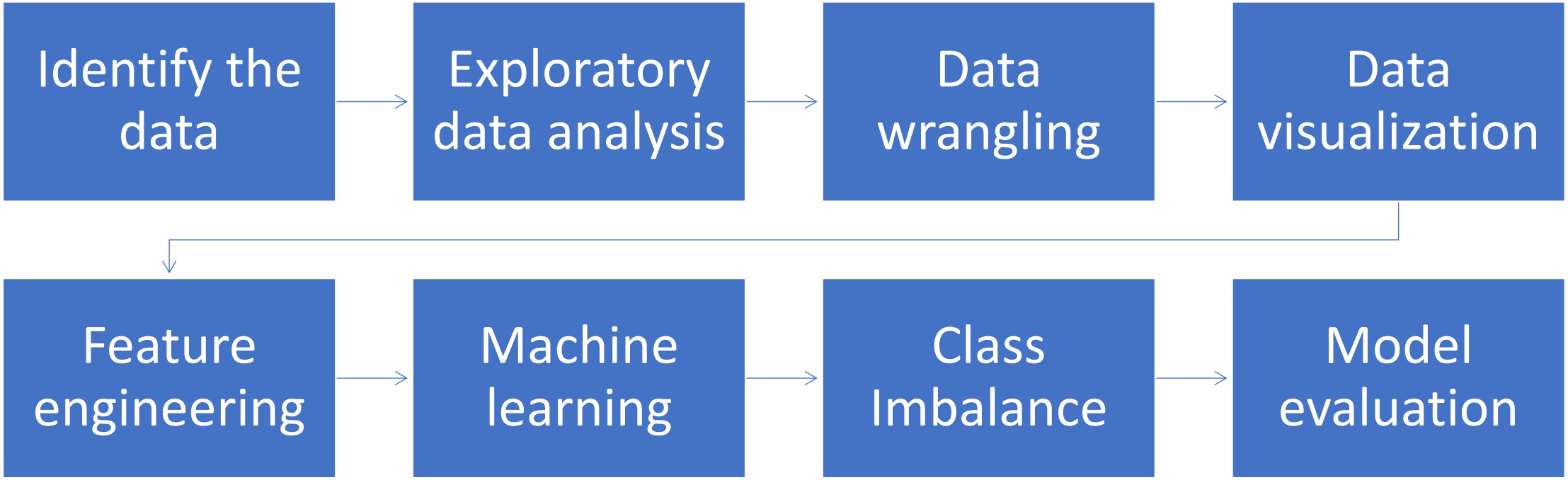
Fintech Companies

- XGBoost
- Logistic Regression
- Scorecard
- Interpretability

Goal

Enable access to credit for people with limited credit history

Steps



Click to access source code

EDA and Data wrangling methods used

307,511 x 122

Target [1,0]

Int64, float64 and object

100,000 missing values

Imputation

Outliers

a) *pd.read_csv:*
b) *df.head():*
   df.rename(columns=str.lower):
c) *pd.set_option:*
d) *df.dtypes:*
e) *df.target.unique():*
f) *df.describe().T*
g) *df_missing=df.isna().sum():*
h) *df[['occupation_type']].fillna(value ='Unknown'):*
i) *outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]*

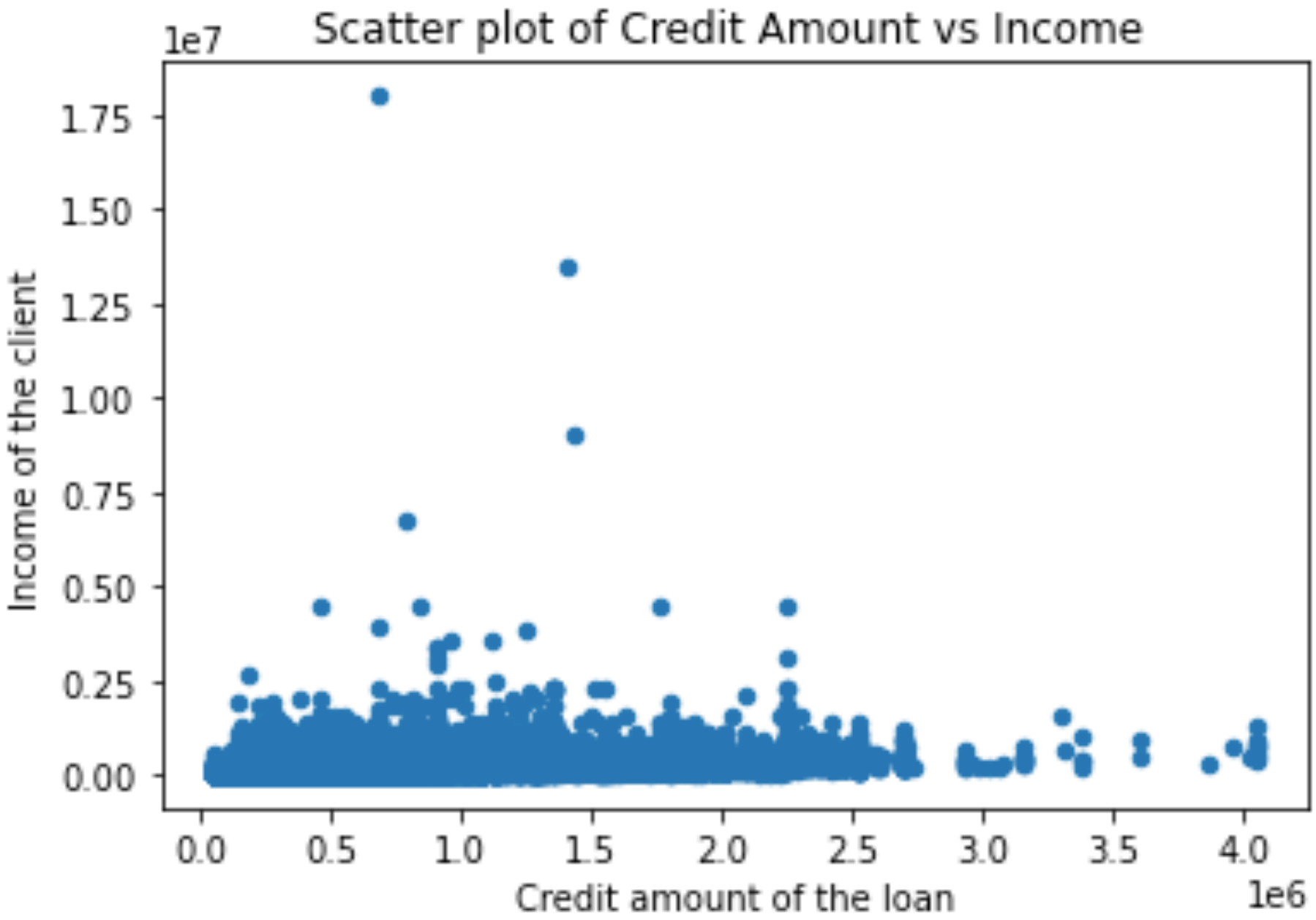
Data Visualization

df.plot(kind='scatter',x==' ',y='')

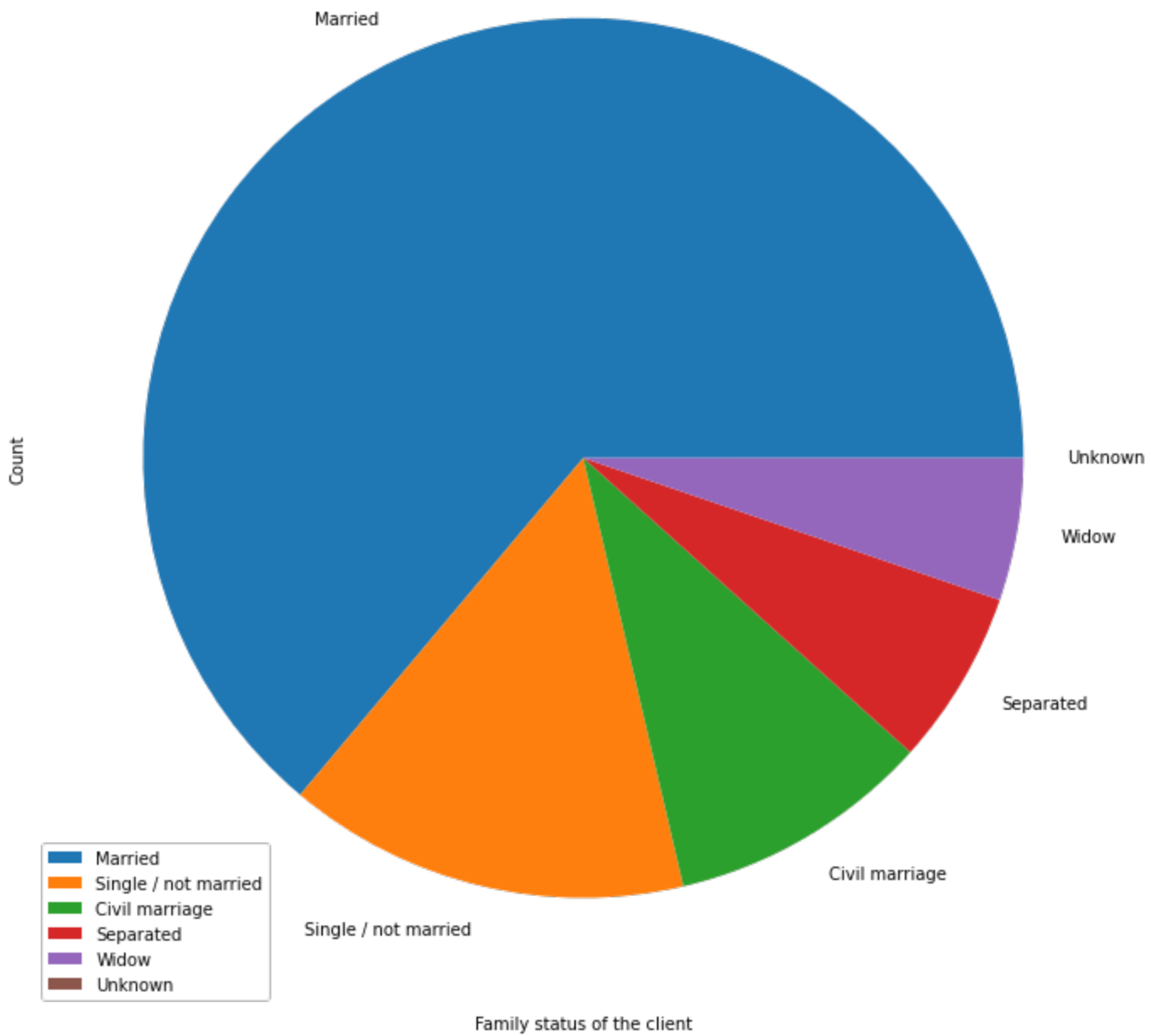
df.plot(kind='pie',legend=True,figsize=(12,12)).

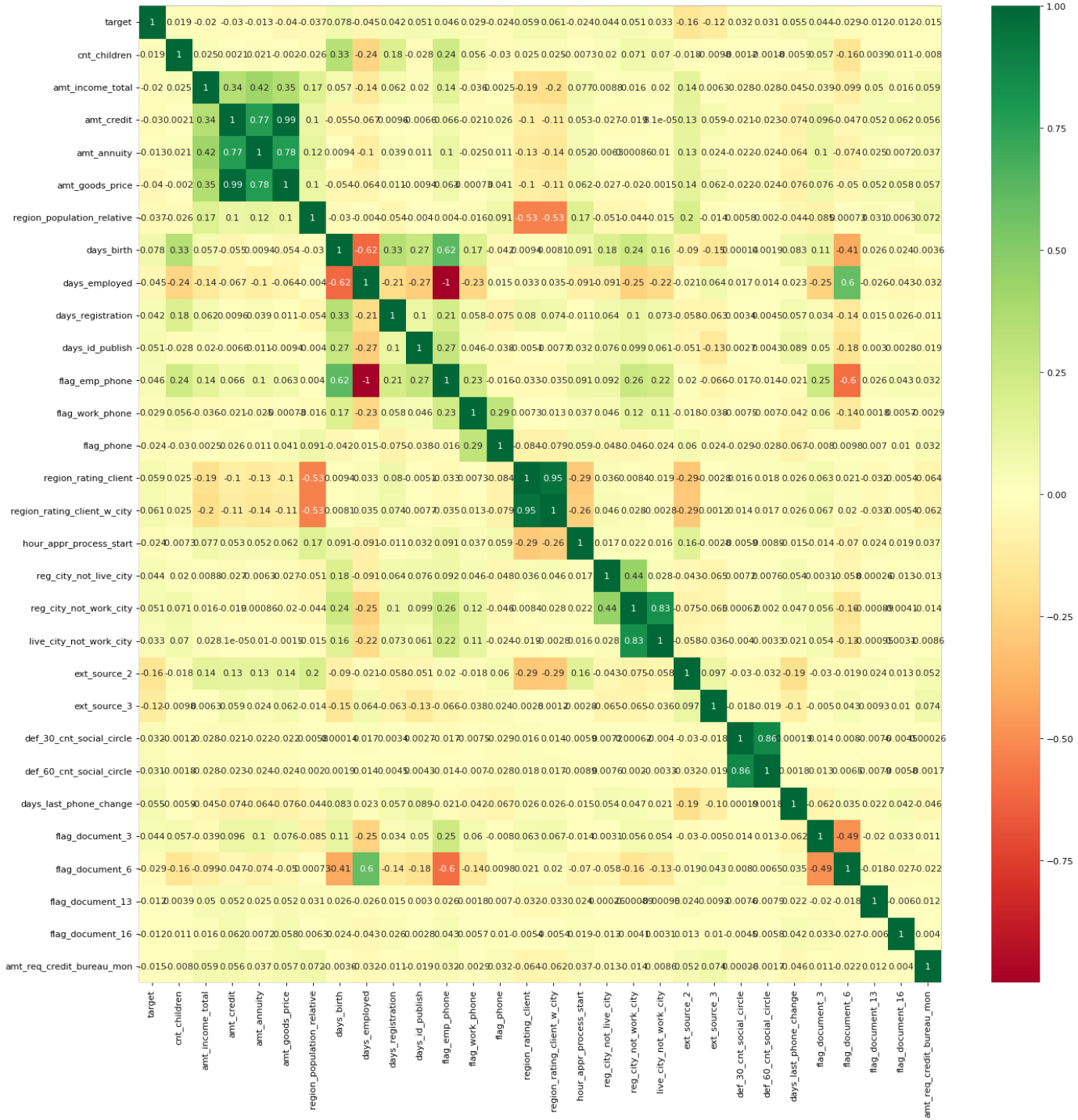
df.plot(kind='bar')

plt.boxplot


Scatter plot of Credit Amount vs Income
1e7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Income of the client
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
Credit amount of the loan

What is the family status of the client?

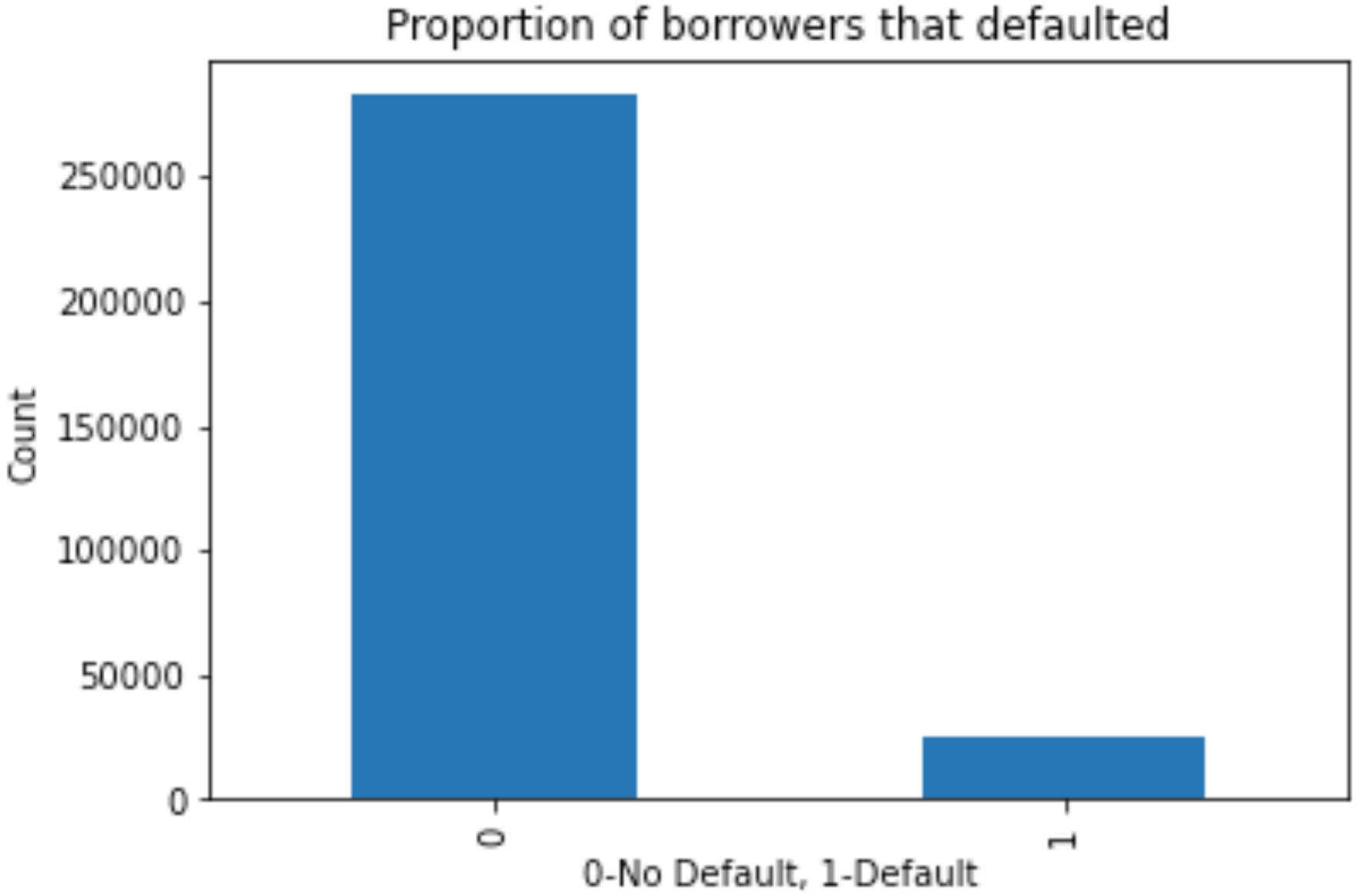



target
cnt_children
amt_income_total
amt_credit
amt_annuity
amt_goods_price
region_population_relative
days_birth
days_employed
days_registration
days_id_publish
flag_emp_phone
flag_work_phone
flag_phone
region_rating_client
region_rating_client_w_city
hour_appr_process_start
reg_city_not_live_city
reg_city_not_work_city
live_city_not_work_city
ext_source_2
ext_source_3
def_30_cnt_social_circle
def_60_cnt_social_circle
days_last_phone_change
flag_document_3
flag_document_6
flag_document_13
flag_document_16
amt_req_credit_bureau_mon
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

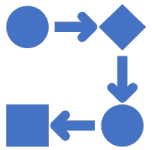
Logistic Regression

```
For the model: LogisticRegression(multi_class='ovr', random_state=0) , the training accuracy
is 0.9191205164059706 and the test accracy is 0.9196936684985854
[[56560     5]
 [ 4934     3]]
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     56565
           1       0.38      0.00      0.00      4937

    accuracy                           0.92     61502
   macro avg       0.65      0.50      0.48     61502
weighted avg       0.88      0.92      0.88     61502
```

Image courtesy: ABC Network

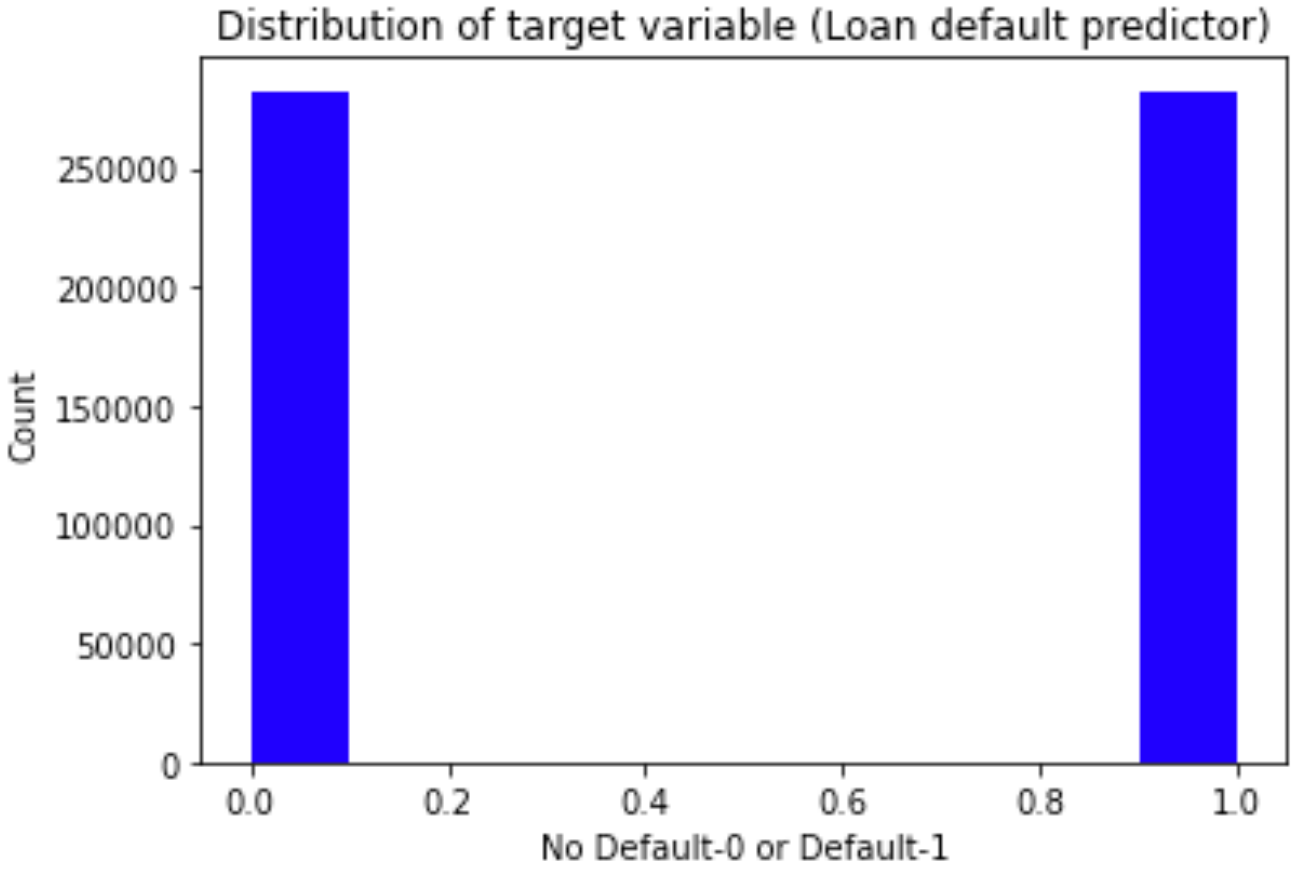

Proportion of borrowers that defaulted
Count
250000
200000
150000
100000
50000
0
0
1
0-No Default, 1-Default

Preparing the data for Machine Learning



Scaling

Class Imbalance
resolution using SMOTE



Extreme gradient boosting with synthetic minority over sampling

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, gamma=None, gpu_id=None, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=None, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=None,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, predictor=None, random_state=None, ...)
Mean cross-validation score: 0.95
K-fold CV average score: 0.95
[[41833   515]
 [ 4122 38336]]
              precision    recall  f1-score   support

           0       0.91      0.99      0.95     42348
           1       0.99      0.90      0.94     42458

    accuracy                           0.95     84806
   macro avg       0.95      0.95      0.95     84806
weighted avg       0.95      0.95      0.95     84806
```

Machine Learning Model Evaluation

Model	Precision	Recall	F1-score	Imbalanced
Logistic Regression	0.38	0.00	0.00	Yes
Stochastic Gradient Descent	0.05	0.00	0.00	Yes
XGBoost	0.55	0.02	0.04	Yes
Stochastic Gradient Descent	0.50	0.00	0.00	No
XGBoost	0.98	0.91	0.94	No
Random Forest	0.94	0.89	0.92	No

Future scope

Deploying the final ML model

continuous variable to generate credit score

Identification of more effective alternative data

Thank You!