Bank Loan Default Prediction

CHRISTOFER BRYAN NATANAEL KOMALA

Problem Statement

Bank loans is one of the major source revenues for banks. The interest charged to the loan applicants is what drives the daily operation of banks. However, bank loans are often associated with risks such as borrowers defaulting on their loans. Banks have collected past data on loan borrowers which include detailed information of each borrower, and they would like to develop a machine learning model to predict if a new borrower is likely to default on their loans or not.

Data Understanding

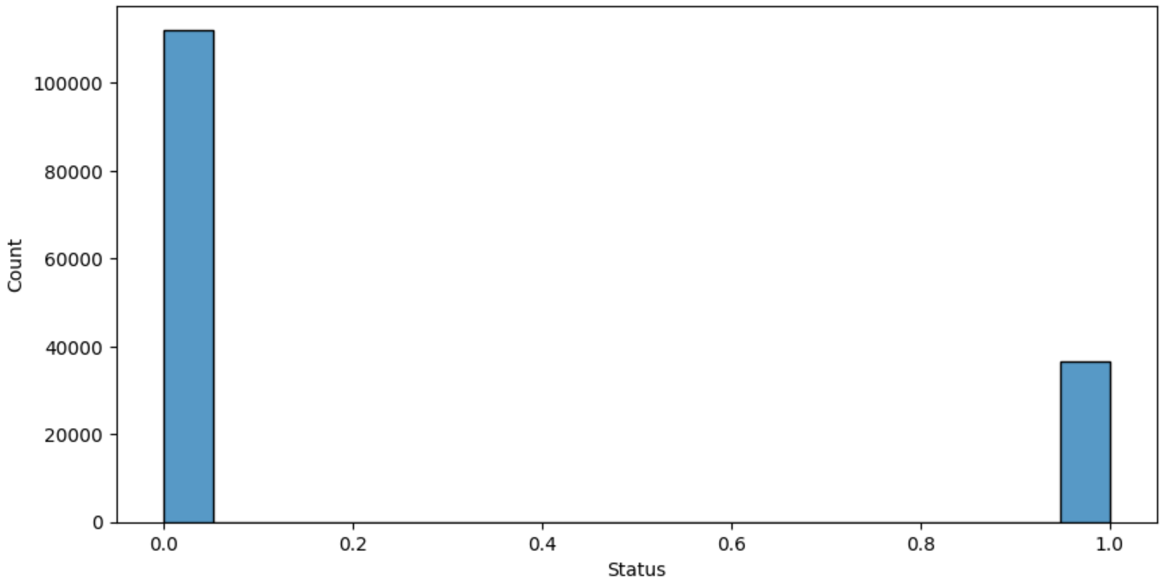
- 148,670 rows
- 34 columns (33 features + 1 target variable)
- Status is the target variable (0 or 1)
- 1 for defaulting applicants and 0 for non-defaulting applicants

```
Data columns (total 34 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   ID                         148670 non-null  int64
 1   year                       148670 non-null  int64
 2   loan_limit                 145326 non-null  object
 3   Gender                     148670 non-null  object
 4   approv_in_adv              147762 non-null  object
 5   loan_type                  148670 non-null  object
 6   loan_purpose               148536 non-null  object
 7   Credit_Worthiness          148670 non-null  object
 8   open_credit                148670 non-null  object
 9   business_or_commercial     148670 non-null  object
 10  loan_amount                148670 non-null  int64
 11  rate_of_interest           112231 non-null  float64
 12  Interest_rate_spread       112031 non-null  float64
 13  Upfront_charges            109028 non-null  float64
 14  term                       148629 non-null  float64
 15  Neg_ammortization          148549 non-null  object
 16  interest_only              148670 non-null  object
 17  lump_sum_payment           148670 non-null  object
 18  property_value             133572 non-null  float64
 19  construction_type          148670 non-null  object
 20  occupancy_type             148670 non-null  object
 21  Secured_by                 148670 non-null  object
 22  total_units                148670 non-null  object
 23  income                     139520 non-null  float64
 24  credit_type                148670 non-null  object
 25  Credit_Score               148670 non-null  int64
 26  co-applicant_credit_type   148670 non-null  object
 27  age                        148470 non-null  object
 28  submission_of_application  148470 non-null  object
 29  LTV                        133572 non-null  float64
 30  Region                     148670 non-null  object
 31  Security_Type              148670 non-null  object
 32  Status                     148670 non-null  int64
 33  dtir1                      124549 non-null  float64
dtypes: float64(8), int64(5), object(21)
```

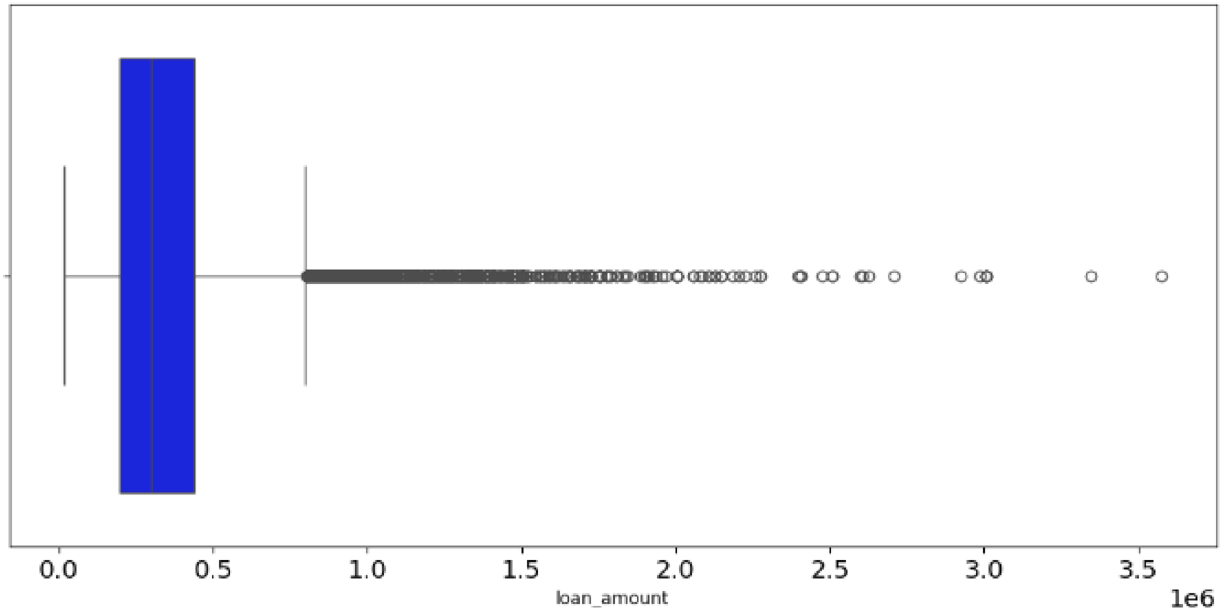
Data Understanding

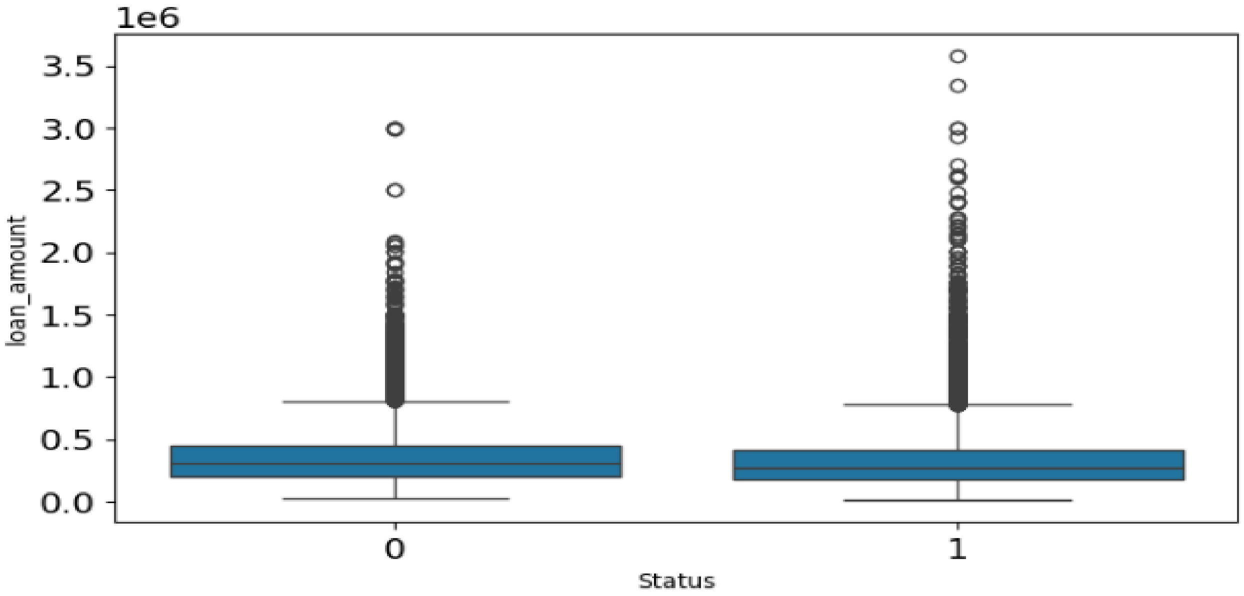
- Some columns contain missing values.
- There are not duplicated records in this data.

```
loan_limit, 3344, 2.2%
approv_in_adv, 908, 0.6%
loan_purpose, 134, 0.1%
rate_of_interest, 36439, 24.5%
Interest_rate_spread, 36639, 24.6%
Upfront_charges, 39642, 26.7%
term, 41, 0.0%
Neg_ammortization, 121, 0.1%
property_value, 15098, 10.2%
income, 9150, 6.2%
age, 200, 0.1%
submission_of_application, 200, 0.1%
LTV, 15098, 10.2%
dtir1, 24121, 16.2%
```



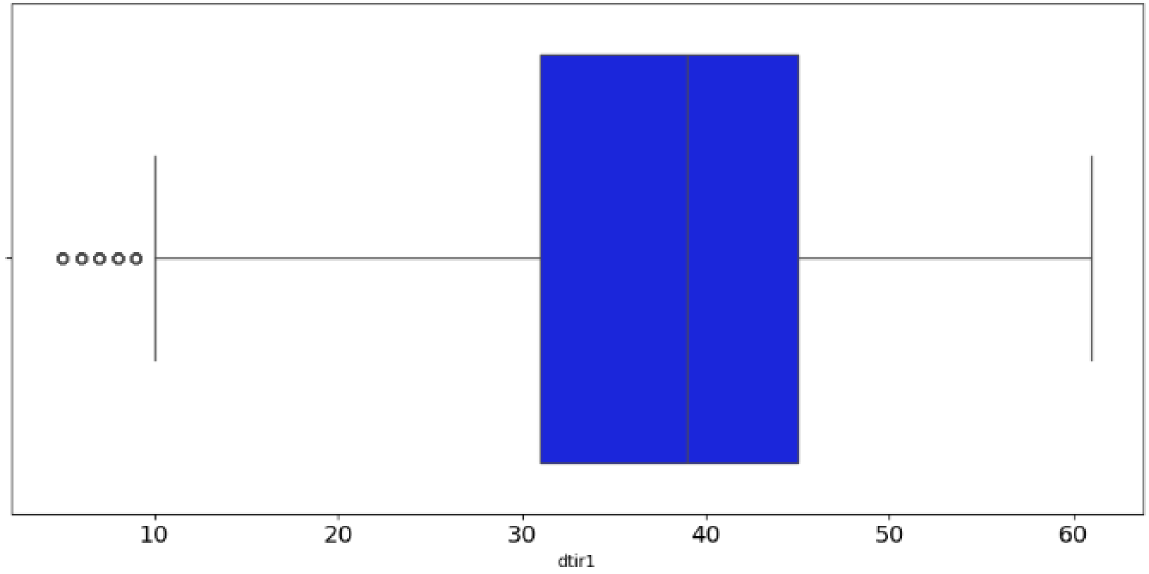
- The Status target column is imbalanced.
- Requires imbalanced data handling.

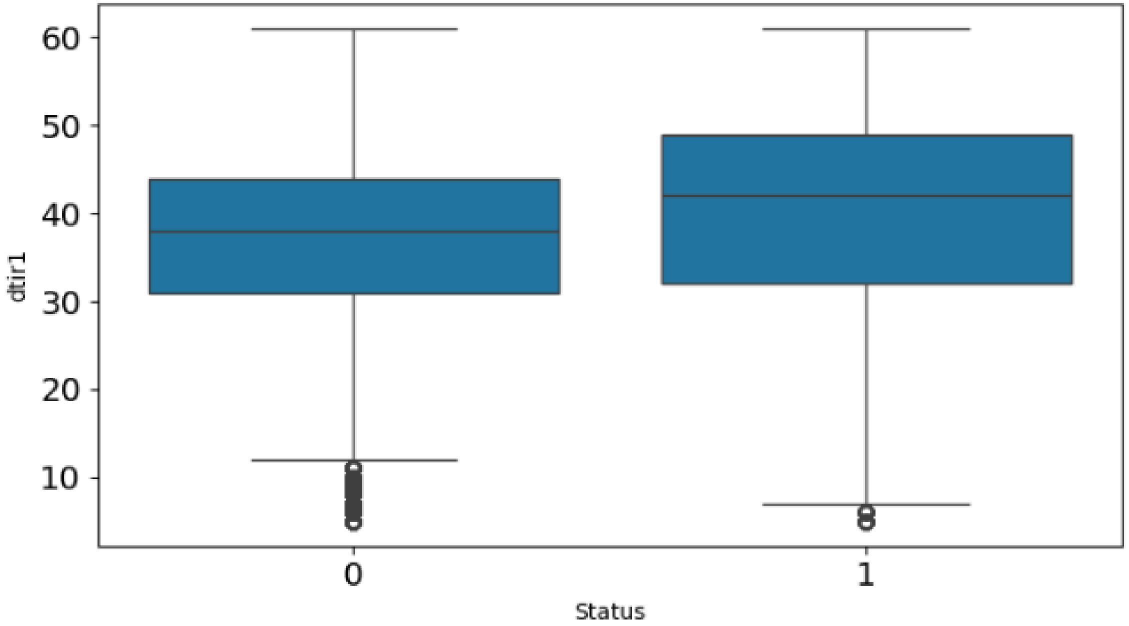




Exploratory Data Analysis

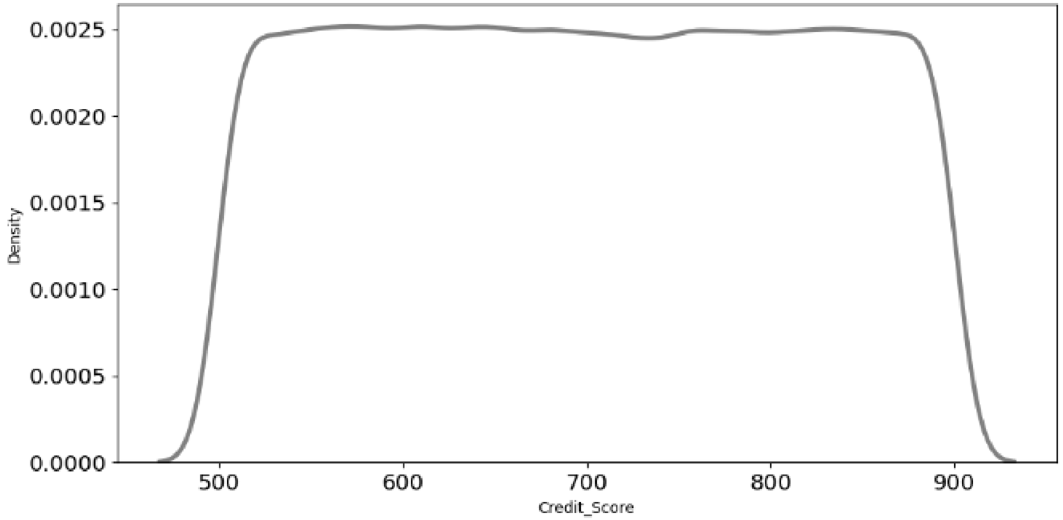
- Loan amount is heavily right skewed. Majority of the applicants applied a loan amount between $0 and $796,500.
- It does not appear to be a determining factor of loan defaults.

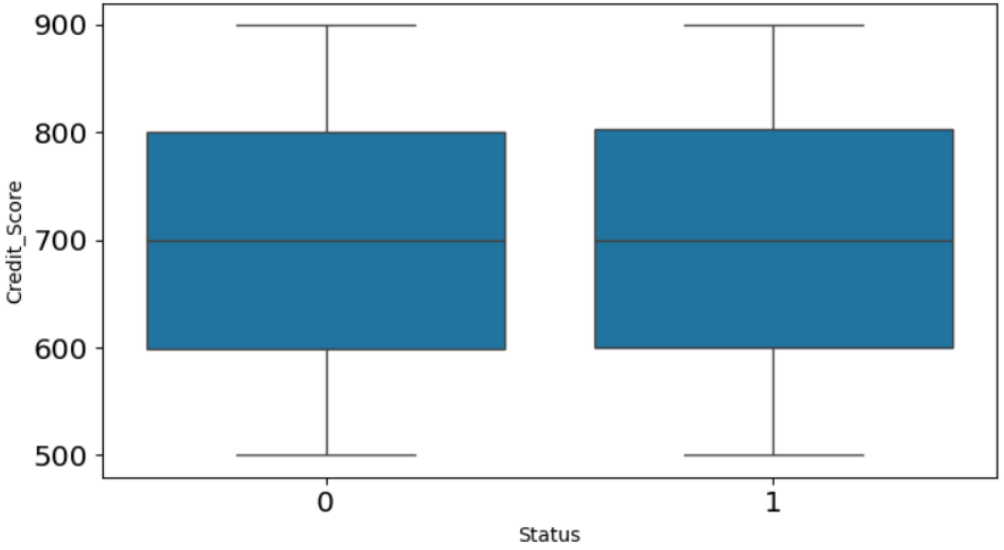




Exploratory Data Analysis

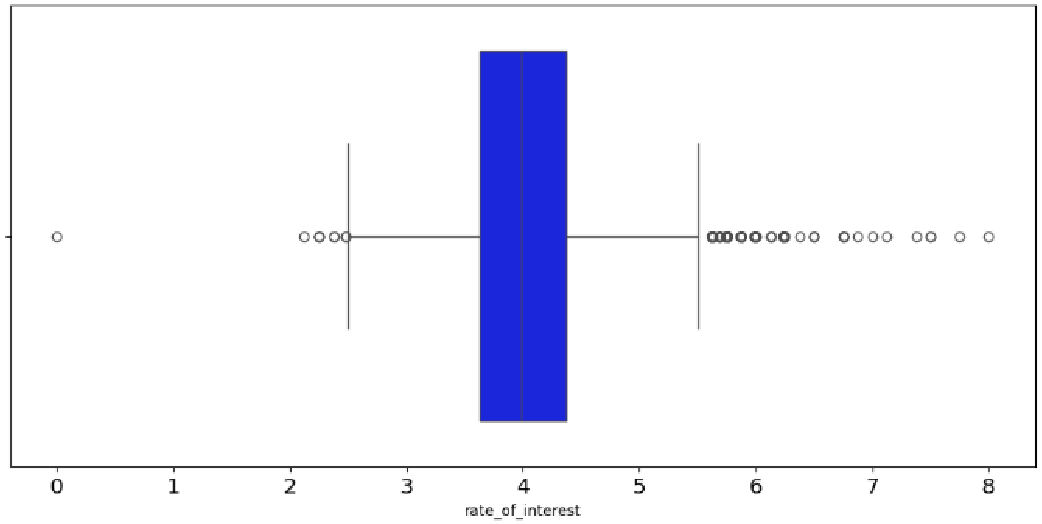
- The debt-to-income ratio is about symmetrical. 50% of the applicants have a ratio of between 31 to 45.
- It appears that applicants with higher debt-to-income ratio have a higher tendency to default on their loans.

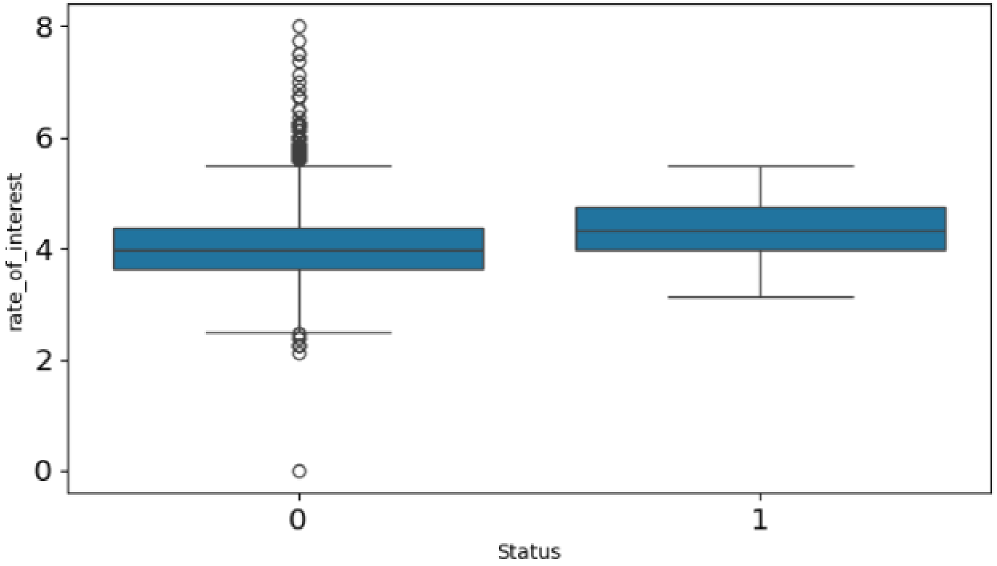




Exploratory Data Analysis

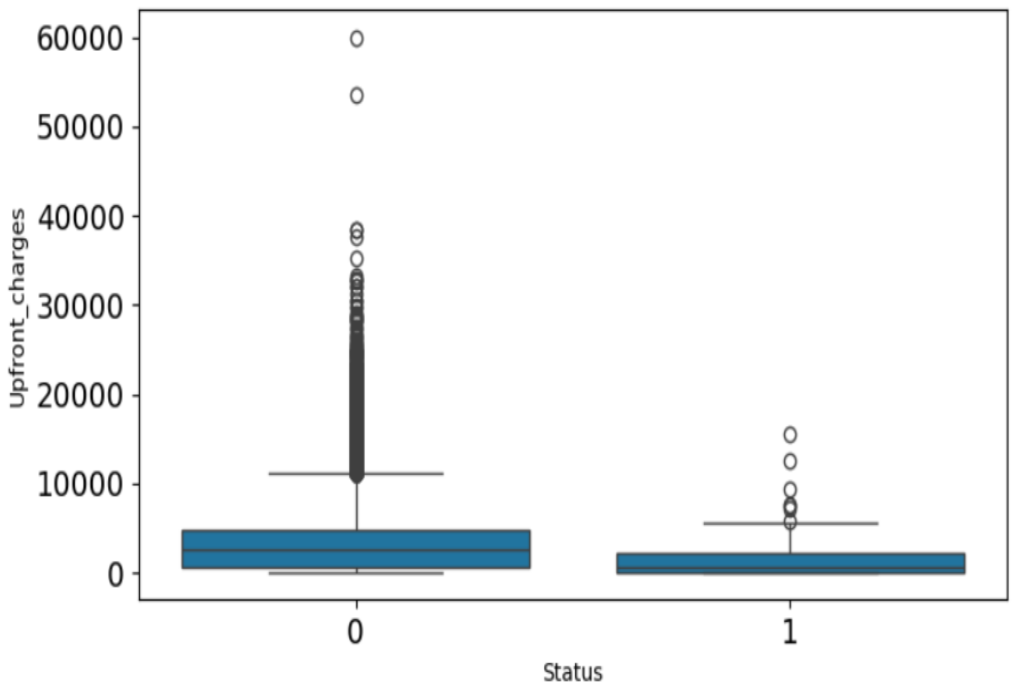
- The credit score has a relatively uniform distribution from 500 to 900.
- Credit scores do not appear to determine whether a loan applicant would default.





Exploratory Data Analysis

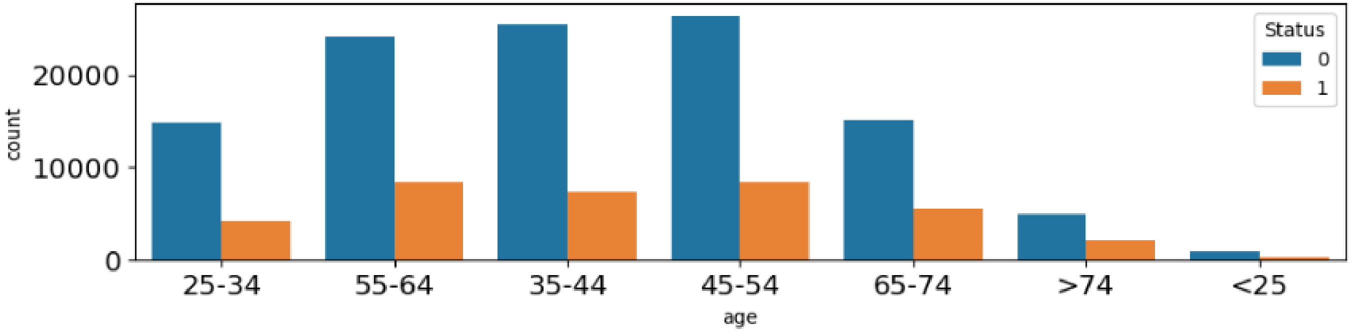
- Applicants charged with higher interest rate are more likely to default on their loans.



Exploratory Data Analysis

- Applicants charged with higher initial loan charges are more likely to not default on their loans.

Exploratory Data Analysis



- Majority of the applicants are of ages 45-54.
- Looking at the applicants of ages 35-44 and 45-54, there is a lower proportion of defaulting applicants.

Exploratory Data Analysis

- All applicants with EQUI credit type default on their loans and there is also a higher number of defaulting applicants who use EQUI compared to other credit types.

Imputing missing values

BEFORE IMPUTING

```
loan_limit, 3344, 2.2%
approv_in_adv, 908, 0.6%
loan_purpose, 134, 0.1%
rate_of_interest, 36439, 24.5%
Interest_rate_spread, 36639, 24.6%
Upfront_charges, 39642, 26.7%
term, 41, 0.0%
Neg_ammortization, 121, 0.1%
property_value, 15098, 10.2%
income, 9150, 6.2%
age, 200, 0.1%
submission_of_application, 200, 0.1%
LTV, 15098, 10.2%
dtir1, 24121, 16.2%
```

AFTER IMPUTING

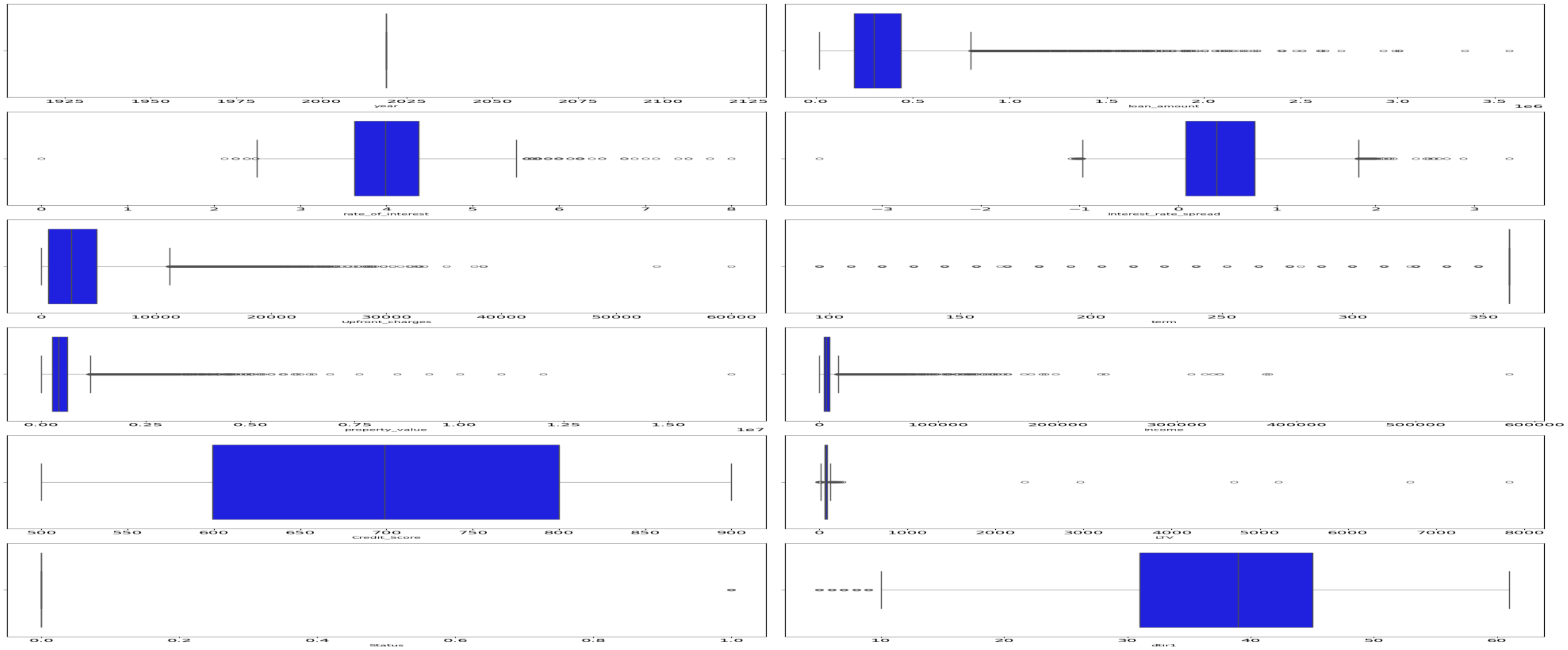
```
loan_limit                   False
Gender                       False
approv_in_adv                False
loan_type                    False
loan_purpose                 False
Credit_Worthiness            False
open_credit                  False
business_or_commercial       False
loan_amount                  False
rate_of_interest             False
Interest_rate_spread         False
Upfront_charges              False
term                         False
Neg_ammortization            False
interest_only                False
lump_sum_payment             False
property_value               False
construction_type            False
occupancy_type               False
Secured_by                   False
total_units                  False
income                       False
credit_type                  False
Credit_Score                 False
co-applicant_credit_type     False
age                          False
submission_of_application    False
LTV                          False
Region                       False
Security_Type                False
dtir1                        False
```

Feature Encoding

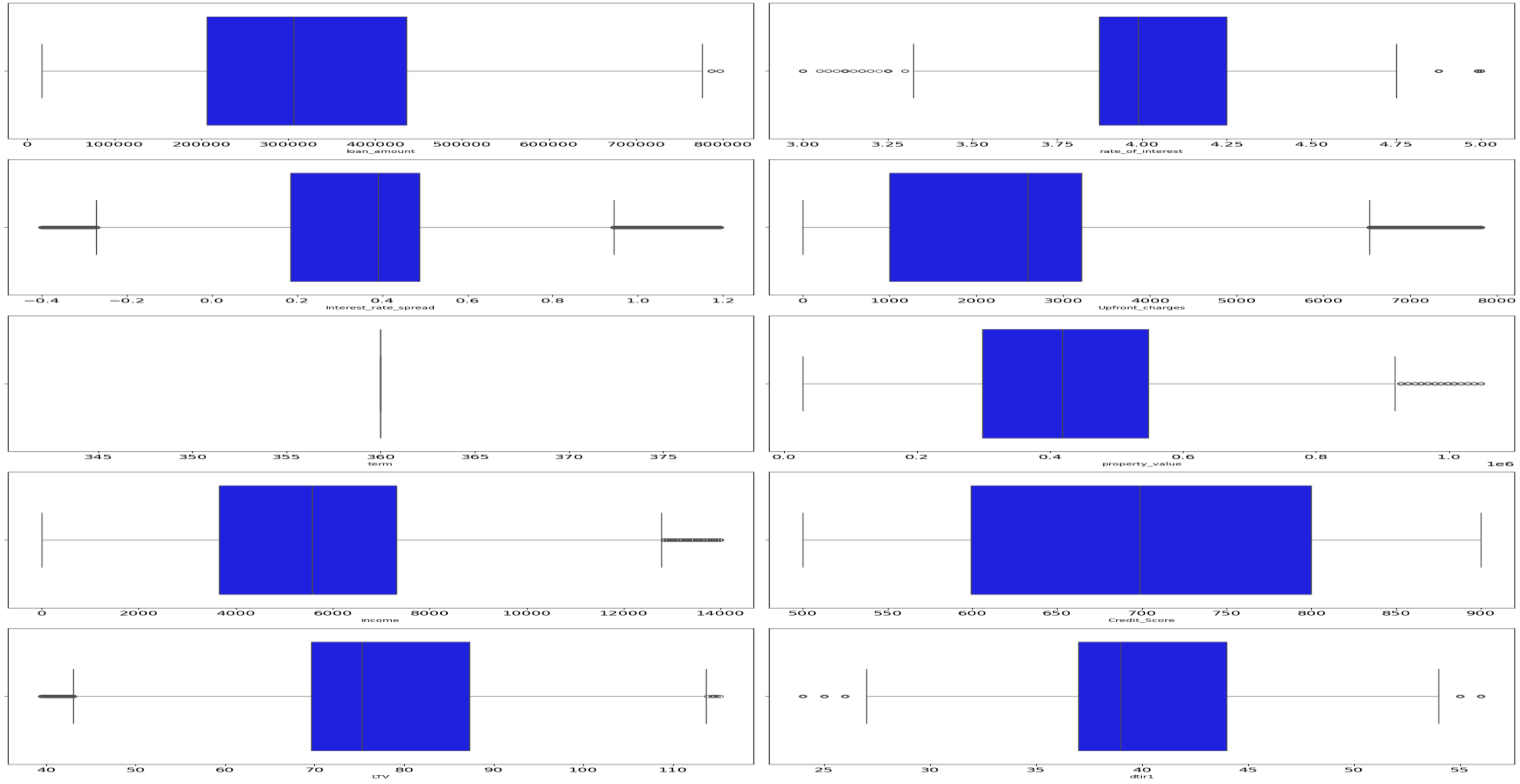
- Convert non-integer categorical columns to integer type by performing label encoding and one hot encoding.

#	Column	Non-Null Count	Dtype
0	loan_limit	104069 non-null	int64
1	approv_in_adv	104069 non-null	int64
2	Credit_Worthiness	104069 non-null	int64
3	open_credit	104069 non-null	int64
4	business_or_commercial	104069 non-null	int64
5	loan_amount	104069 non-null	int64
6	rate_of_interest	104069 non-null	float64
7	Interest_rate_spread	104069 non-null	float64
8	Upfront_charges	104069 non-null	float64
9	term	104069 non-null	float64
10	Neg_ammortization	104069 non-null	int64
11	interest_only	104069 non-null	int64
12	lump_sum_payment	104069 non-null	int64
13	property_value	104069 non-null	float64
14	construction_type	104069 non-null	int64
15	Secured_by	104069 non-null	int64
16	income	104069 non-null	float64
17	Credit_Score	104069 non-null	int64
18	co-applicant_credit_type	104069 non-null	int64
19	submission_of_application	104069 non-null	int64
20	LTV	104069 non-null	float64
21	Security_Type	104069 non-null	int64
22	dtir1	104069 non-null	float64
23	Gender_Joint	104069 non-null	float64
24	Gender_Male	104069 non-null	float64
25	Gender_Sex Not Available	104069 non-null	float64
26	loan_type_type2	104069 non-null	float64
27	loan_type_type3	104069 non-null	float64
28	loan_purpose_p2	104069 non-null	float64
29	loan_purpose_p3	104069 non-null	float64
30	loan_purpose_p4	104069 non-null	float64
31	occupancy_type_pr	104069 non-null	float64
32	occupancy_type_sr	104069 non-null	float64
33	total_units_2U	104069 non-null	float64
34	total_units_3U	104069 non-null	float64
35	total_units_4U	104069 non-null	float64
36	credit_type_CRIF	104069 non-null	float64
37	credit_type_EQUI	104069 non-null	float64
38	credit_type_EXP	104069 non-null	float64
39	age_35-44	104069 non-null	float64
40	age_45-54	104069 non-null	float64
41	age_55-64	104069 non-null	float64
42	age_65-74	104069 non-null	float64

Outlier Handling



Outlier Handling



Feature Scaling

BEFORE STANDARDIZING

	year	loan_amount	rate_of_interest	Interest_rate_spread	Upfront_charges
count	148670.0	1.486700e+05	112231.000000	112031.000000	109028.000000
mean	2019.0	3.311177e+05	4.045476	0.441656	3224.996127
std	0.0	1.839093e+05	0.561391	0.513043	3251.121510
min	2019.0	1.650000e+04	0.000000	-3.638000	0.000000
25%	2019.0	1.965000e+05	3.625000	0.076000	581.490000

AFTER STANDARDIZING

	loan_limit	approv_in_adv	Credit_Worthiness	open_credit	loan_amount	rate_of_interest	Interest_rate_spread	Upfront_charges
count	5.608500e+04	5.608500e+04	5.608500e+04	5.608500e+04	5.608500e+04	5.608500e+04	5.608500e+04	5.608500e+04
mean	3.243273e-17	-7.664765e-18	1.393594e-18	5.447684e-18	-1.469608e-17	-3.234658e-15	-1.508882e-16	1.266903e-17
std	1.000009e+00	1.000009e+00	1.000009e+00	1.000009e+00	1.000009e+00	1.000009e+00	1.000009e+00	1.000009e+00

Feature Selection

Dropped features:

1. “business_or_commercial”
2. “property_value”
3. “term”
4. “construction_type”
5. “Secured_by”
6. “Security_type”
7. “loan_type_type2”

- Select features using VIF (variance inflation factor). Features with high VIF imply strong multicollinearity between them and are dropped.

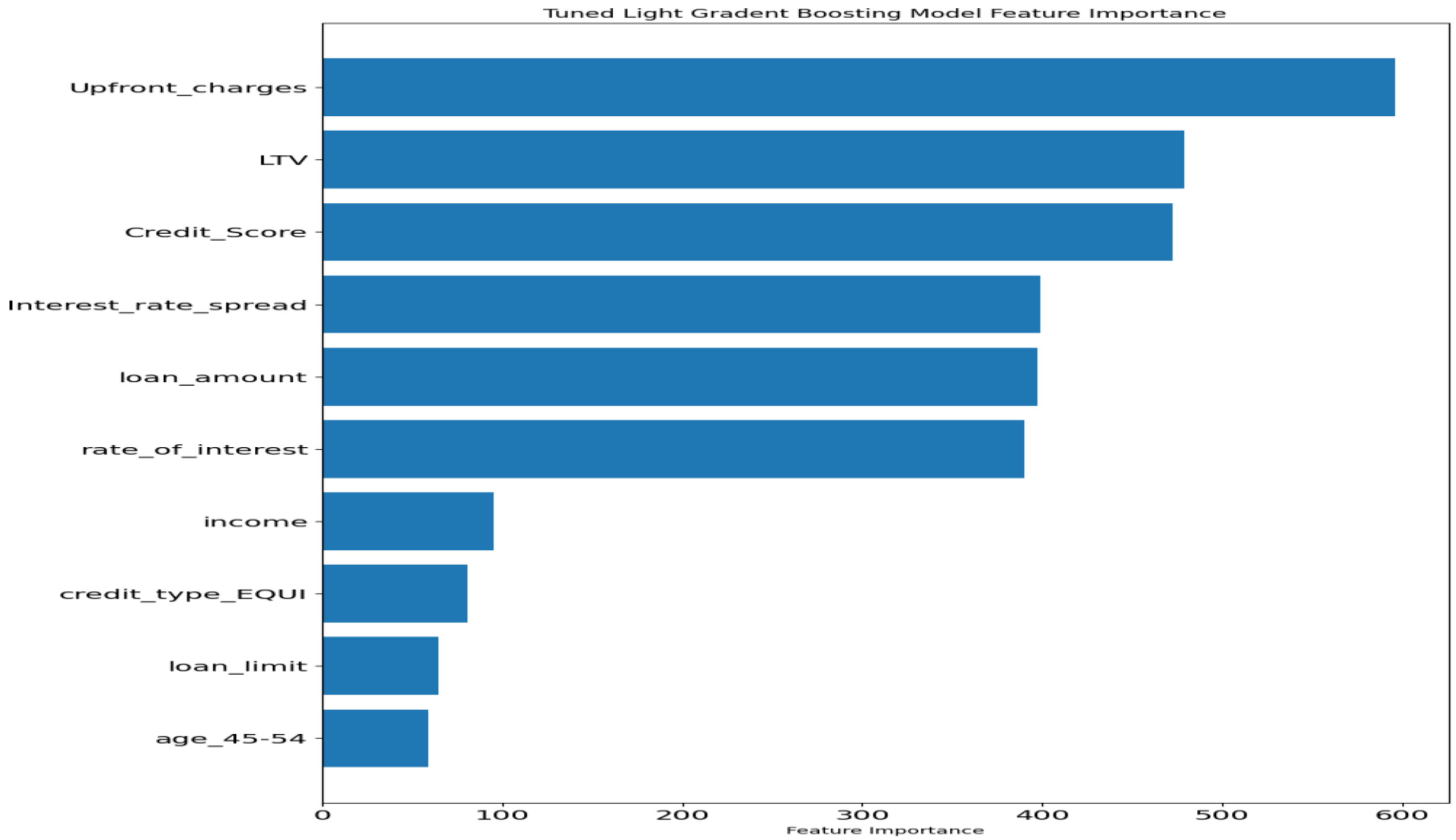
Modelling

Model	Recall (Train)	Recall (Test)	F1 score (Train)	F1 score (Test)	ROC AUC score (Train)	ROC AUC score (Test)
Logistic Regression	0.726	0.650	0.791	0.665	0.808	0.775
Tuned Logistic Regression	0.726	0.650	0.791	0.665	0.808	0.775
K-nearest neighbors	0.986	0.849	0.927	0.716	0.923	0.840
Random Forest	1.000	1.000	1.000	0.999	1.000	0.999
Tuned Random Forest	1.000	1.000	1.000	1.000	1.000	1.000
LGBM	1.000	1.000	1.000	0.999	1.000	0.999
Tuned LGBM	1.000	1.000	1.000	1.000	1.000	1.000

The dataset is split into 70% training data and 30% testing data before data preprocessing to ensure no testing data leakage.

In the end, we select K-nearest neighbors as our model as it does not appear to be overfitting unlike random forest and LGBM. It also performs better (higher recall) than logistic regressions which appears to be underfitting.

Feature Importance



Recommendations

- Focus more on applicants with a lower debt-to-income ratio as they are more likely to repay their loans (not default).
- Lower the interest rate for loan borrowers as higher interest rates tend to make applicants to default more.
- Focus more on the two age groups 35-44 and 45-54 years old as they comprise majority of the total applicants and have lower proportion of defaulting applicants.
- Avoid approving loans for applicants using credit type of EQUI.
- Increase upfront charges so that only people who can really afford to repay their loans are approved to borrow from the bank. These people are more likely to successfully repay their loans.