Data Mining Project

Home Credit Default Risk

KAMAL PATEL

ABOUT KAGGLE COMPETITION

- **Home Credit Group** is committed to financial inclusion, offering a positive borrowing experience to those with limited credit history.
- Home Credit assesses repayment abilities beyond traditional methods for a more inclusive approach.
- Home Credit optimize data usage to predict repayment accurately to prevent rejections and empower clients with tailored loans for success.

PROBLEM DESCRIPTION

- Many people struggle to get loans due to insufficient credit histories

- Aim to make use of alternative data (e.g. previous info, credit card info) to predict repayment ability

- Ensure customer with capability to repay loans are **NOT** rejected

OBJECTIVE

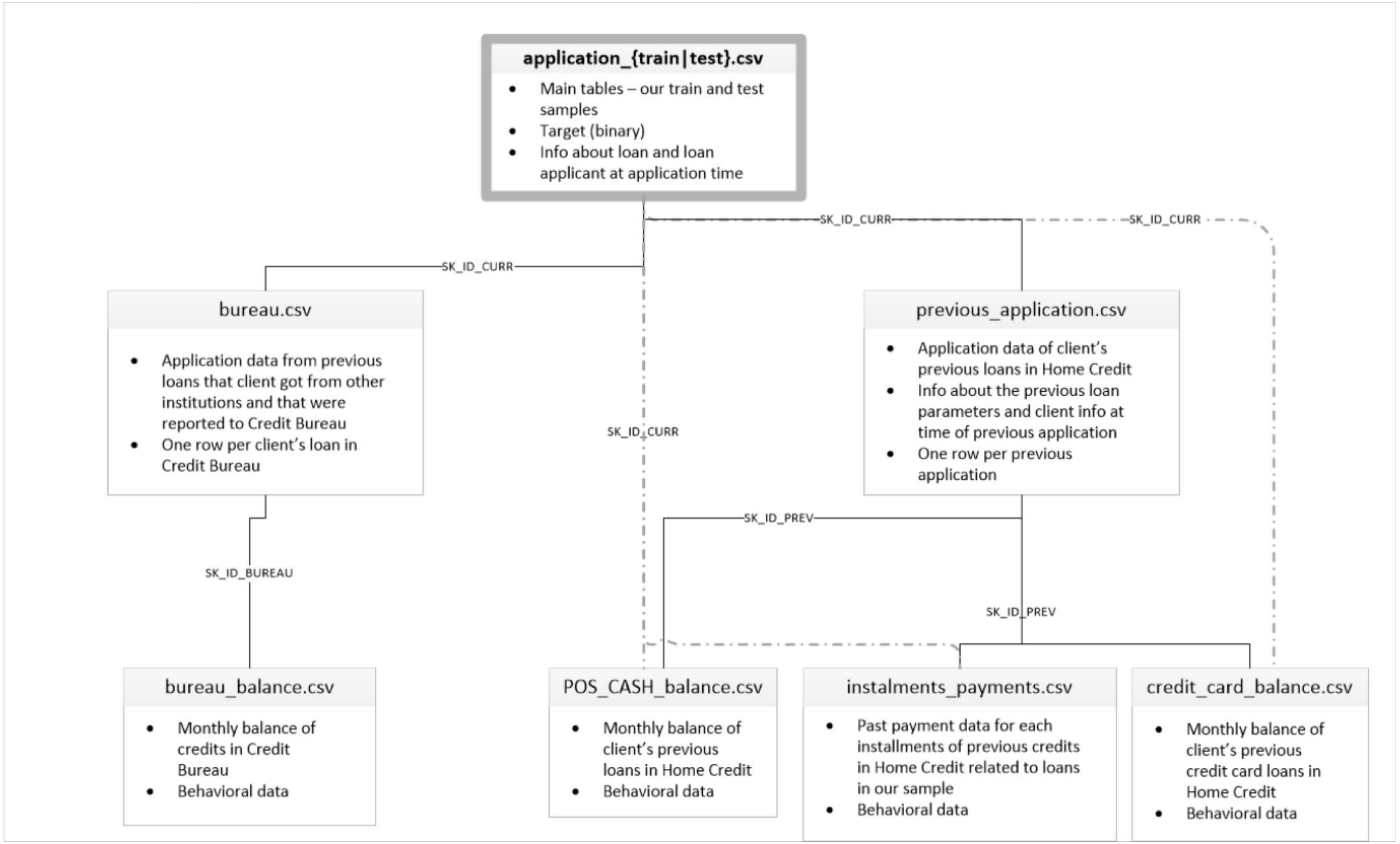
Predict if each application is NOT capable of repaying a loan

Classify into 2 Classes:

Negative – 0: Able to repay a loan ✓

Positive – 1: Unable to repay a loan ✗

DATA DESCRIPTION



DATA TERMINOLOGY

application.csv: Description of few column in the dataset

Column	Description	Values
NAME_CONTRACT_TYPE	Identification of loan	{Cash, Revolving}
AMT_INCOME_TOTAL	Income of the client	numerical value eg. 12000.65
AMT_CREDIT	Credit amount of the loan	numerical value eg. 16734.6
NAME_FAMILY_STATUS	Family status of client	{single, marries, not_married}
NAME_EDUCATION_TYPE	Level of client's education	{Secondary, higher, Incomplete}
DAYS_BIRTH	Client's age in days	numerical values eg.-9461, -16765
NAME_INCOME_TYPE	Clients income type	{business, working, pensioner}

DATA MATRIX

300,000 Rows and 122 features (106 Numerical, 16 Categorical).

Each row is one loan application.

Below is the first few rows and columns of dataset for **application_train.csv.**

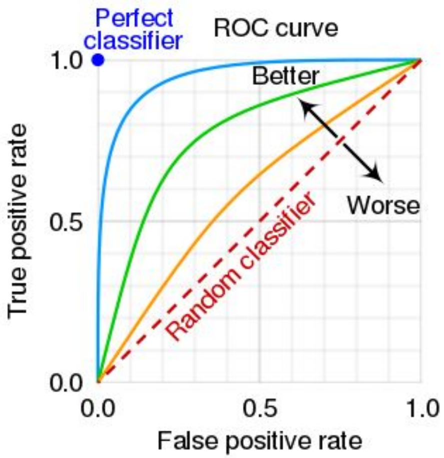
SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALITY	….	AMT_INCOME_TOTAL
100002	1	Cash loans	M	N	Y		202500.0
100003	0	Cash loans	F	N	N	….	270000.0
100004	0	Revolving loans	M	Y	Y	….	67500.0
10006	0	Cash loans	F	N	Y	….	135000.0
…							

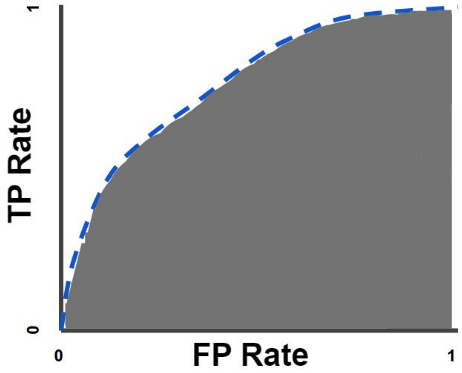
To Predict

EVALUATION METRICS

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

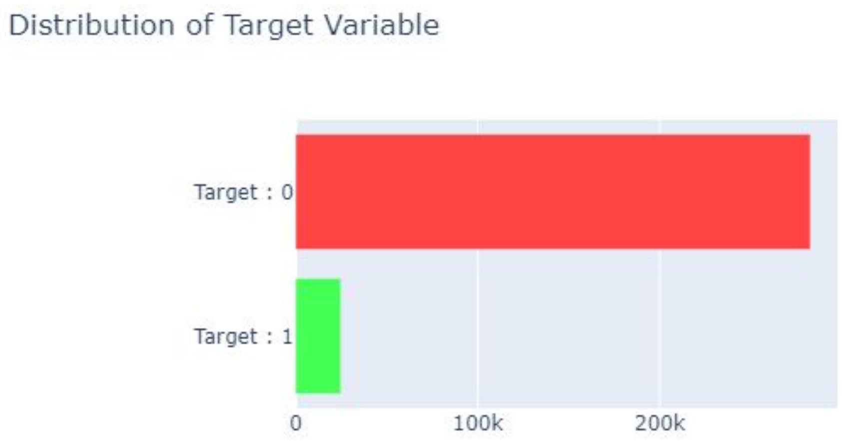
AUC is the probability that the model ranks a random positive example more highly than a random negative example





EXPLORATORY DATA ANALYSIS

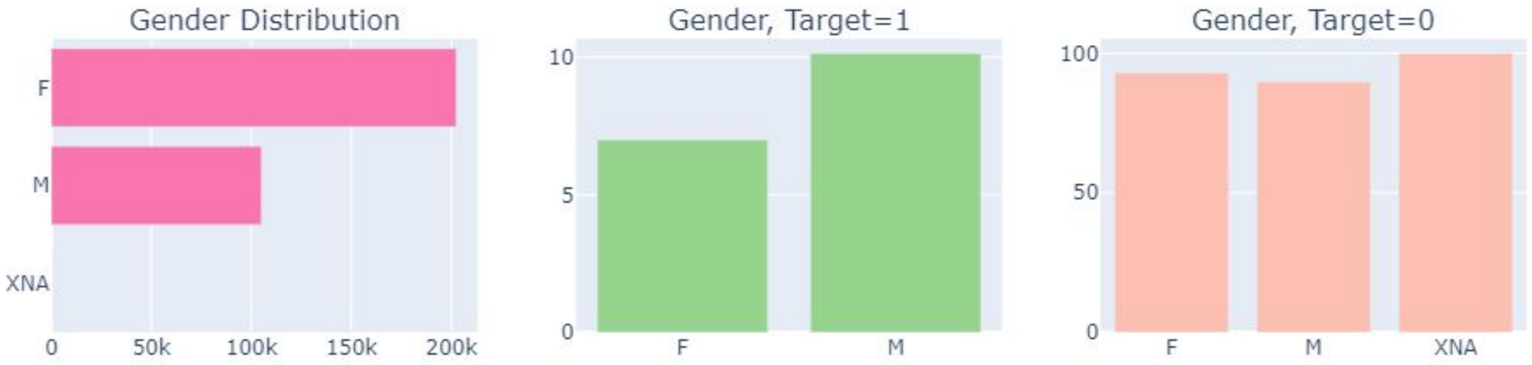
TARGET VARIABLE



0: Able to repay a loan

1: Unable to repay a loan

CLIENT'S GENDER



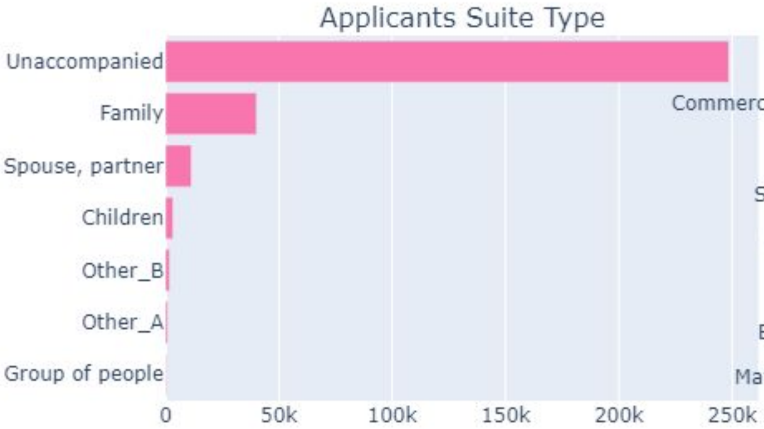
There are about 202,448 loan applications filed by females in contrast to about 105,059 applications filed by males.

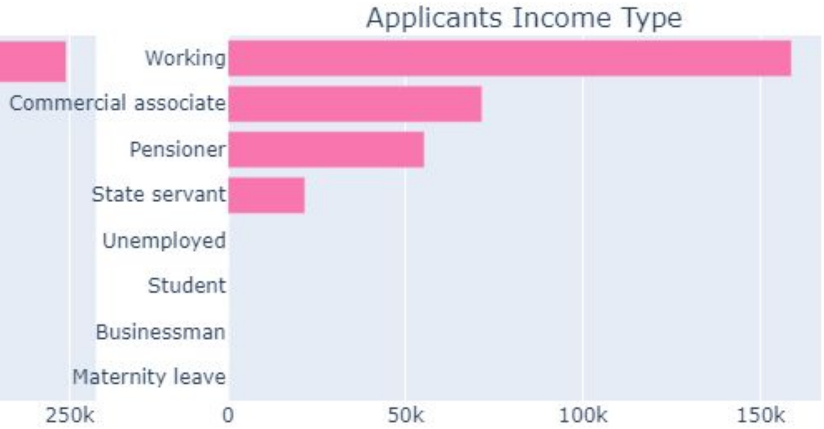
10% of men had the problems in paying the loan or making installments compared to women applicants about 7%.

FAMILY STATUS

10% of men had the problems in paying the loan or making installments compared to women applicants about 7%.

SUITE TYPE & INCOME TYPE





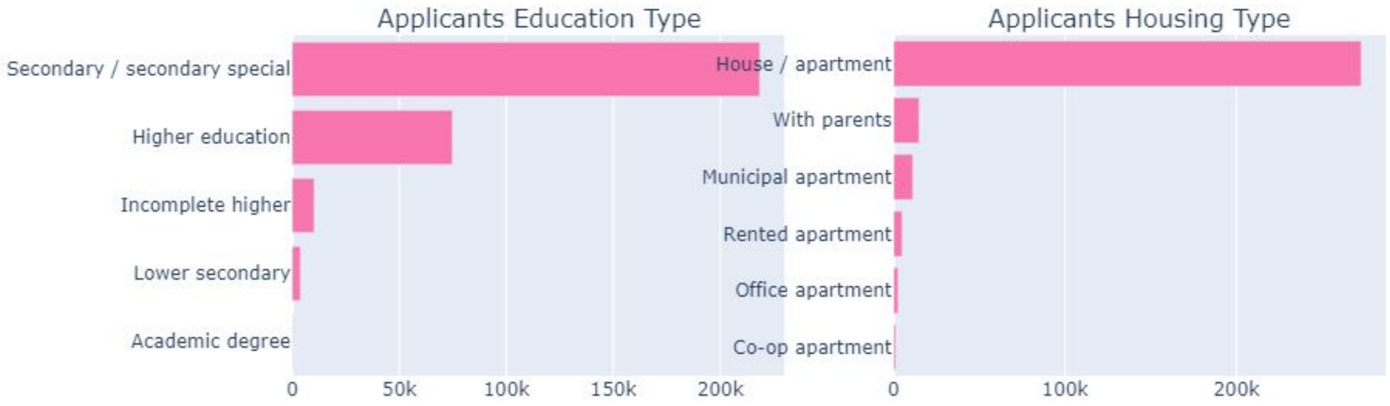
Suite Type

- Unaccompanied (about 248K applicants)
- Family (about 40K applicants)
- Spouse, partner (about 11K applicants)

Income type

- Working Class (158K),
- Commercial Associate (71K)
- Pensiner (55K)

EDUCATION & HOUSING TYPE



Education Type

- 218K loan application filed by people having secondary education.
- 75K by people with Higher Education.

Housing Type

- Applicants living in House / apartments has the highest number of loan applications equal to 272K

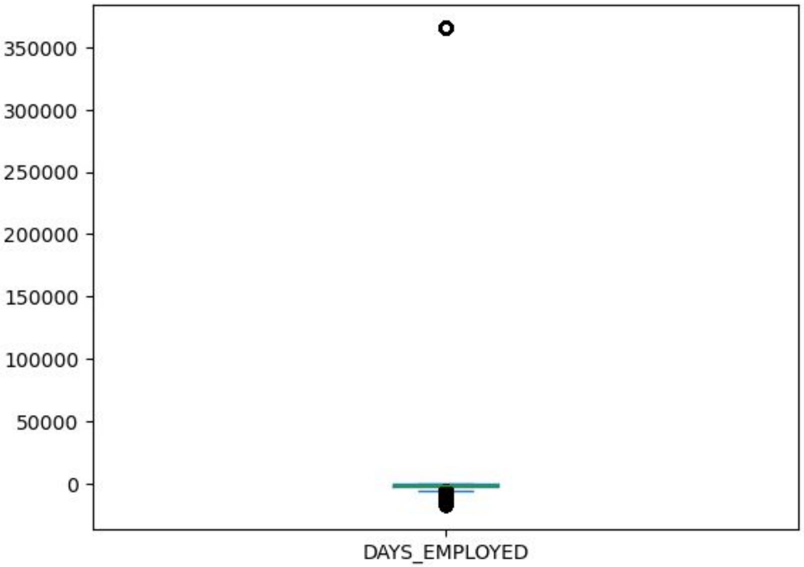
Data Preprocessing

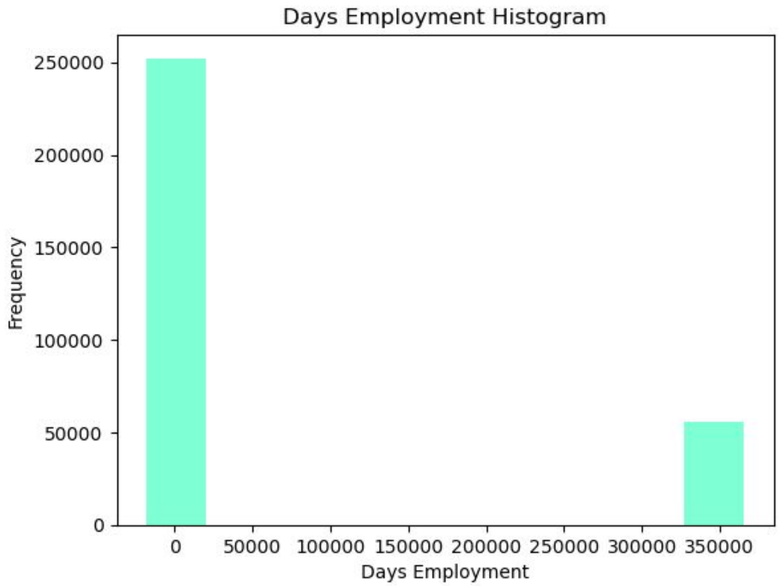
Anomaly Detection	Find values that deviate from the normal range
Missing Values	Find Missing value in both categorical and numerical columns and perform preprocessing.
Label Encoding	Encode categorical columns with label or one-hot encoding based on number of categories.
Normalization	Transform the column in dataset to same scale.

Anomaly Detection

strange value: 365243, it could mean empty values or some errors but deeper analysis is required.

The days of employment is recorded relative to the current loan application date and therefore should be negative like the majority of other values in that column.





Missing values

```
Your selected dataframe has 122 columns.
There are 72 columns that have missing values.
```

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_MODE	214865	69.9
COMMONAREA_AVG	214865	69.9
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
...	...	...
DEF_30_CNT_SOCIAL_CIRCLE	1021	0.3
OBS_60_CNT_SOCIAL_CIRCLE	1021	0.3
DEF_60_CNT_SOCIAL_CIRCLE	1021	0.3
EXT_SOURCE_2	660	0.2
AMT_GOODS_PRICE	278	0.1

67 rows × 2 columns

Impute numerical column with median values

For categorical value I used mode imputation

Label & ONE HOT ENCODING

Label Encoding

- For columns with 2 or fewer categories (binary values)

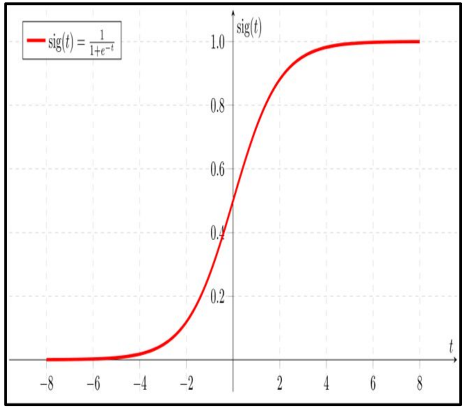
One Hot Encoding

- For columns with more than 3 categories.

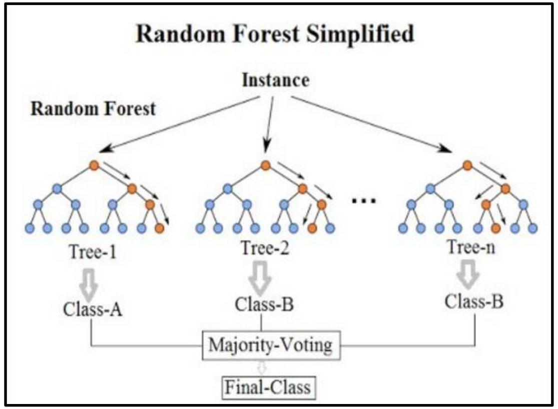
	Column	Number_of_Categories
0	NAME_CONTRACT_TYPE	2
1	CODE_GENDER	2
2	FLAG_OWN_CAR	2
3	FLAG_OWN_REALTY	2
4	NAME_TYPE_SUITE	7
5	NAME_INCOME_TYPE	7
6	NAME_EDUCATION_TYPE	5
7	NAME_FAMILY_STATUS	5
8	NAME_HOUSING_TYPE	6
9	OCCUPATION_TYPE	18
10	WEEKDAY_APPR_PROCESS_START	7
11	ORGANIZATION_TYPE	57
12	FONDKAPREMONT_MODE	4
13	HOUSETYPE_MODE	3
14	WALLSMATERIAL_MODE	7
15	EMERGENCYSTATE_MODE	2

ALGORITHMS USED

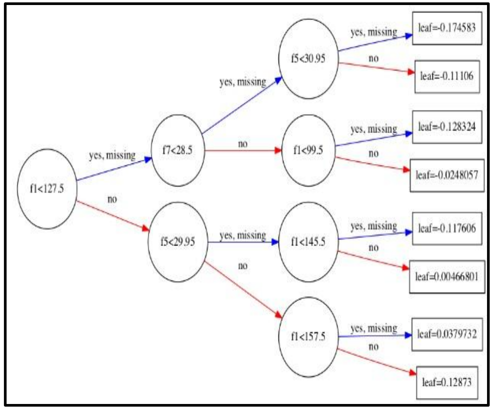
Logistic Regression



Random Forest



Boosting Methods



MODEL COMPARISON

Scores from kaggle submission

Models	AUC
Logistic Regression (Base)	0.63
Random Forest	0.694
XGBoost	0.74
Light GBM	0.731

XGBoost performs the best with 0.74 auc score without feature engineering (future scope)

Total participants in competition 8,373

Top 100 participants scored in between 0.79 and 0.8.

1st Place 0.80570

TUNING MODEL

- As the dataset is very large and tuning parameter using RandomizedSearchCV and GridSearchCV are time consuming.
- So I decided to use AutoML library like pycaret and h20.ai to get best parameters values for XGBoost and LightGBM.

After Model Tuning

Models	ROC AUC
XGBoost	0.73606
Light GBM	0.74367

Best model after tuning is LightGBM with 0.74367

KAGGLE PERFORMANCE

BEFORE

lgbm.csv Complete (after deadline) · 2d ago · lgbm	0.73173	0.73136
xgb.csv Complete (after deadline) · 2d ago · xgboost	0.73898	0.74093
rf.csv Complete (after deadline) · 2d ago · rf	0.69349	0.69458
logreg_baseline.csv Complete (after deadline) · 2d ago · logreg	0.61527	0.63042

AFTER

Submission and Description	Private Score	Public Score
tune_lgbm.csv Complete (after deadline) · 2m ago · tune lgbm	0.74293	0.74367
tuned_xgb.csv Complete (after deadline) · 3m ago · tune xgb	0.73502	0.73606

CONCLUSION

- LightGBM is currently the best chosen model, Until further work.
- Applying domain knowledge for feature engineering with given different dataset could improve the performance of the model.
- Advanced techniques like SMOTE could be deployed to handle the class imbalance problems.
- Performing GridSearchCV for better hyperparameter tuning

THANK YOU.