Loan Prediction

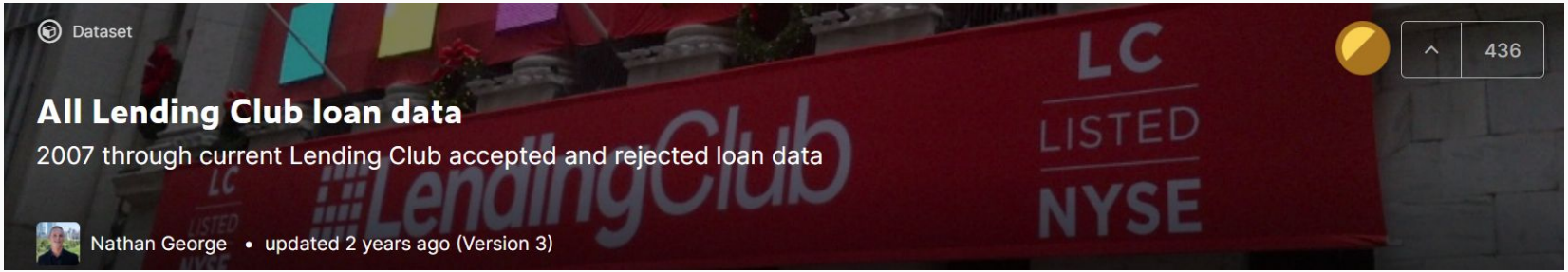
Yitian Cauthen

Problem Statement

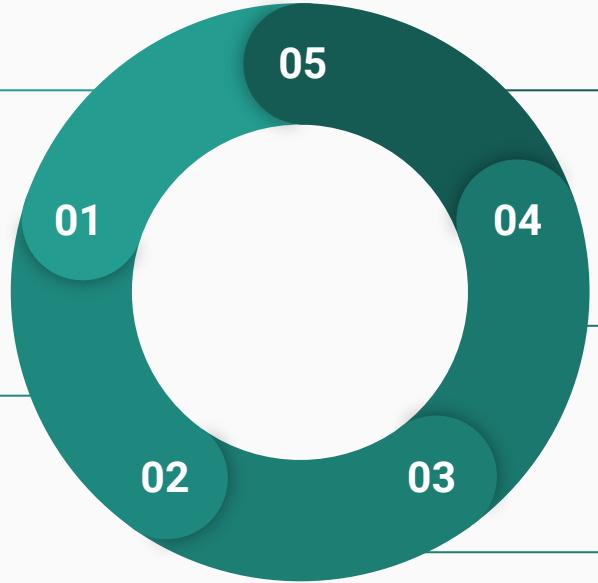
How can we use classification algorithms to predict the probability that a given person will default on a personal loan given their attributes i.e. income, current credit rating, years of employment, loan grade, e.t.c.

Source of Data

- Taken from Kaggle, the Lending Club dataset contains 2,000,000 observations with 151 features of accepted loans between 2007 and 2018



Methodology and Approach



Data Cleaning/Wrangling

EDA Part 1

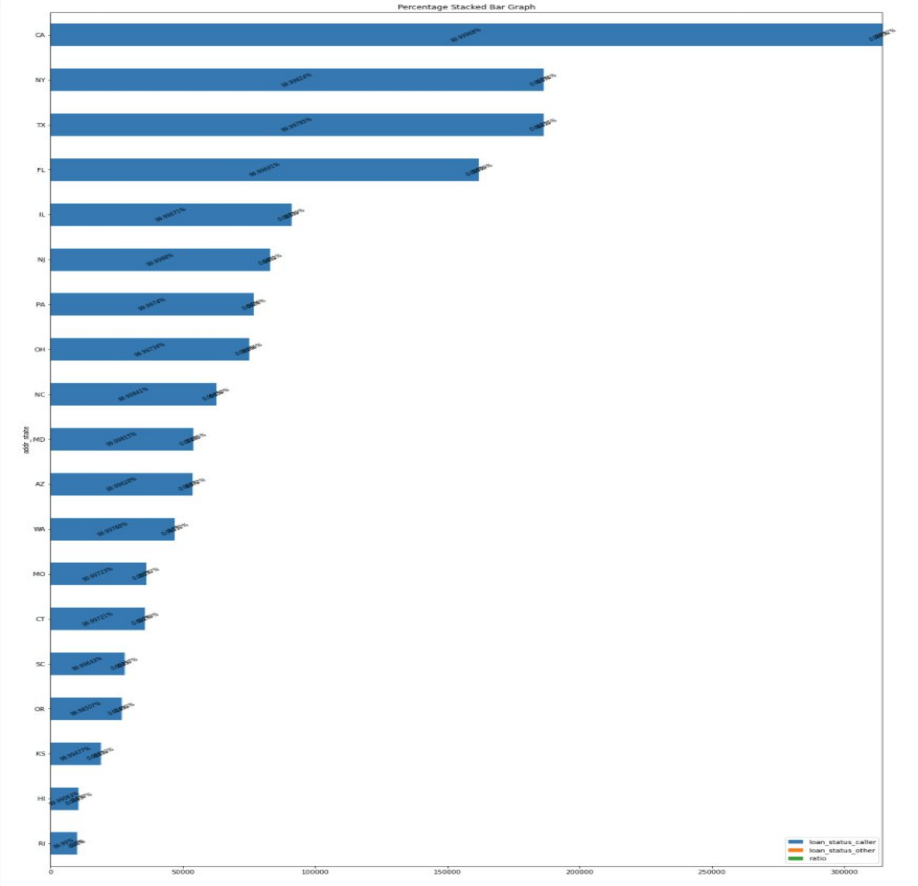
EDA Part 2

Preprocessing/Training and Baseline Modeling

Classification Modeling and Evaluation

Exploratory Data Analysis: Loans By State

	addr_state	loan_status_caller	loan_status_other	ratio
0	RI	10004	1.00000	0.00002
1	HI	10667	1.00000	0.00000
2	KS	19108	1.00000	0.00003
3	OR	26785	4.00000	0.00001
4	SC	28002	1.00000	0.00009
5	CT	35784	1.00000	0.00001
6	MO	36083	1.00000	0.00005
7	WA	47059	1.00000	0.00002
8	AZ	53775	2.00000	0.00003
9	MD	54007	1.00000	0.00002
10	NC	62729	1.00000	0.00001
11	OH	75130	2.00000	0.00001
12	PA	76937	2.00000	0.00001
13	NJ	83131	1.00000	0.00004
14	IL	91170	3.00000	0.00001
15	FL	161986	5.00000	0.00010
16	TX	186331	4.00000	0.00004
17	NY	186382	7.00000	0.00001
18	CA	314532	1.00000	0.00002



Modeling

Logistic Regression

Random Forest

XGBoost

LightGBM

SMOTE

	Model	Accuracy	Precision	Recall	F1	Support
0	lr_train_0	0.98790	1.00000	0.98000	0.99000	400000
1	lr_train_1	0.98790	0.98000	1.00000	0.99000	400000
2	lr_test_0	0.97520	1.00000	0.98000	0.99000	100000
3	lr_test_1	0.97520	0.00000	0.62000	0.00000	8
4	rf_test_0	0.99990	1.00000	1.00000	1.00000	100000
5	rf_test_1	0.99990	0.00000	0.00000	0.00000	8
6	xgb_train_0	1.00000	1.00000	1.00000	1.00000	400000
7	xgb_train_1	1.00000	1.00000	1.00000	1.00000	400000
8	xgb_test_0	1.00000	1.00000	1.00000	1.00000	100000
9	xgb_test_1	0.00000	0.00000	0.00000	0.00000	8
10	lgbm_train_0	1.00000	1.00000	1.00000	1.00000	400000
11	lgbm_train_1	1.00000	1.00000	1.00000	1.00000	400000
12	lgbm_test_0	0.99990	1.00000	1.00000	1.00000	100000
13	lgbm_test_1	0.99990	0.00000	0.00000	0.00000	8

ADASYN

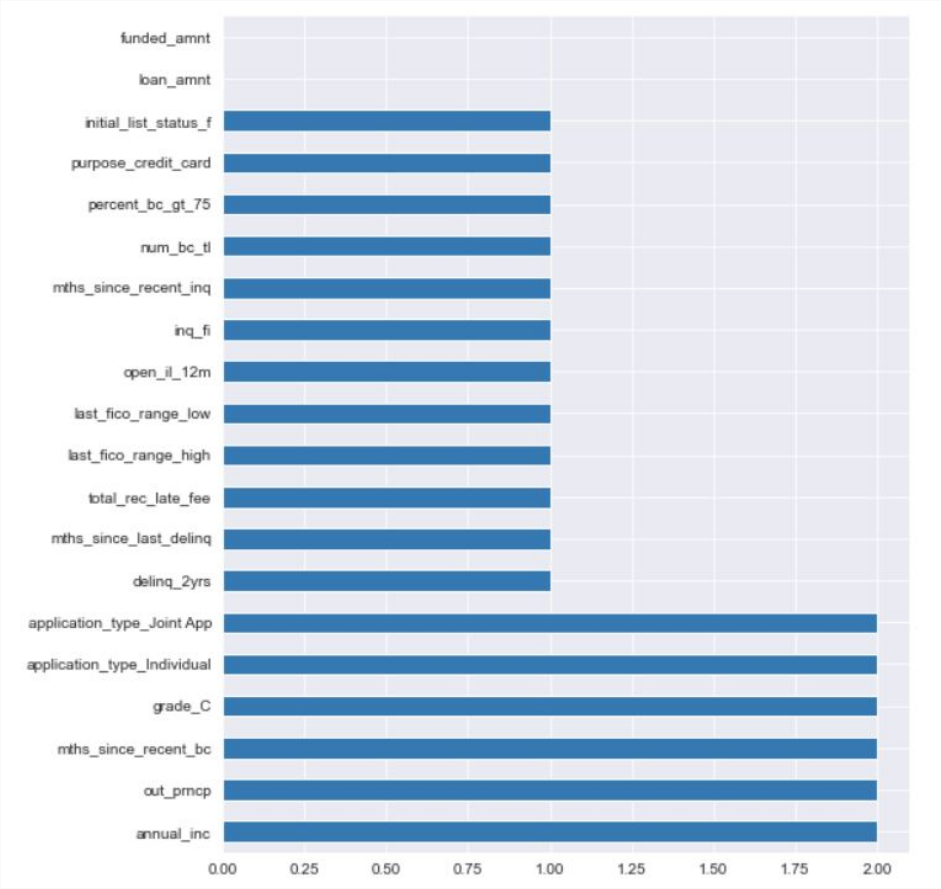
	Model	Accuracy	Precision	Recall	F1	Support
0	lr_train_0	0.98790	1.00000	0.98000	0.99000	400000
1	lr_train_1	0.98790	0.98000	1.00000	0.99000	400000
2	lr_test_0	0.97520	1.00000	0.98000	0.99000	100000
3	lr_test_1	0.97520	0.00000	0.62000	0.00000	8
4	rf_test_0	0.99990	1.00000	1.00000	1.00000	100000
5	rf_test_1	0.99990	0.00000	0.00000	0.00000	8
6	xgb_train_0	1.00000	1.00000	1.00000	1.00000	400000
7	xgb_train_1	1.00000	1.00000	1.00000	1.00000	400000
8	xgb_test_0	0.99990	1.00000	1.00000	1.00000	100000
9	xgb_test_1	0.99990	0.00000	0.00000	0.00000	8
10	lgbm_train_0	1.00000	1.00000	1.00000	1.00000	400000
11	lgbm_train_1	1.00000	1.00000	1.00000	1.00000	400000
12	lgbm_test_0	0.99990	1.00000	1.00000	1.00000	100000
13	lgbm_test_1	0.99990	0.00000	0.00000	0.00000	8

LGBM w/ Hyperparameter Optimization (SMOTE)

Hyperparameters and Best Estimates:
'colsample_bytree': 0.9234
'min_child_samples': 399
'min_child_weight': 0.1
'num_leaves': 13
reg_alpha': 2
'reg_lambda': 5
'subsample': 0.855

Model	Accuracy	Precision	Recall	F1	Support
lgbm_tuned_tr_0	0.99	1.00	0.99	1.00	400000
lgbm_tuned_tr_1	0.99	1.00	1.00	1.00	400000
lgbm_tuned_tst_0	0.99	1.00	0.99	1.00	100000
lgbm_tuned_tst_1	0.99	0.00	0.10	0.00	20

Feature Importance



Conclusion

- Annual income, grade, remaining outstanding principal, months since recent balance, and application type had the highest feature importance and may affect the probability of default in loan risk assessment
- Logistic Regression and LGBM w/ hyperparameter optimization performed the best compared to Random Forest and XGBoost in predicting probability of default
- Further work should be done and models can be enhanced

Sources:

Lending Club Image:

https://mura.heap.io/sites/default/cache/file/6810B6B9-A5E6-4268-9C0A669BBD59FE20.png

Lending Club Dataset:

https://www.kaggle.com/wordsforthewise/lending-club