Credit Analysis with Machine Learning Models

November 2023

Project overview

Identify which machine learning model is best for accurately predicting whether a customer will default and determine which features are most relevant for classifying customers

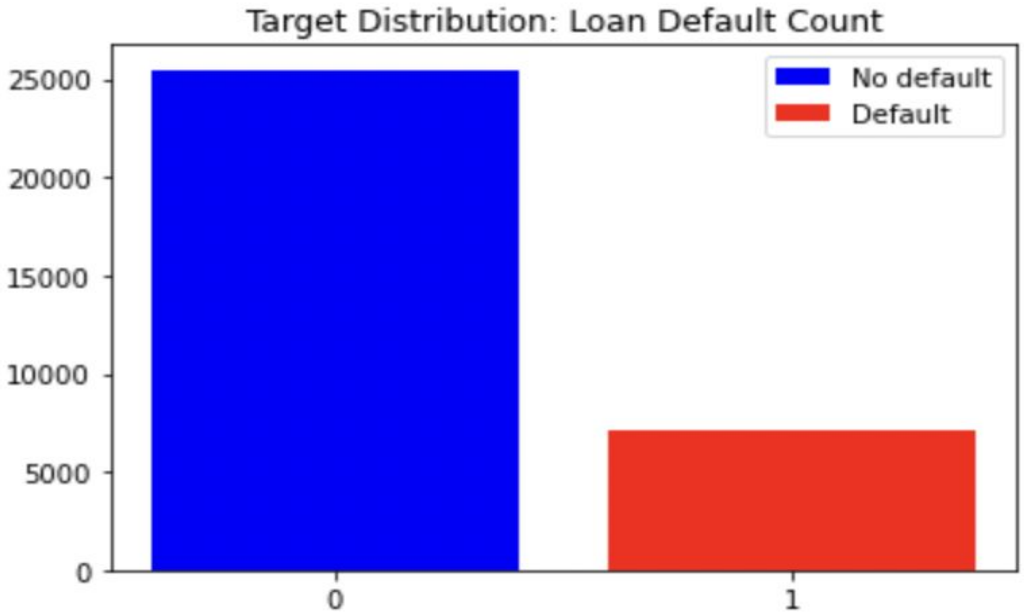
The original dataset

Our data comes from information recorded by the credit bureau

person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income
123.0	PERSONAL	D	35000	16.02	1	0.59
5.0	EDUCATION	B	1000	11.14	0	0.10
1.0	MEDICAL	C	5500	12.87	1	0.57
4.0	MEDICAL	C	35000	15.23	1	0.53
8.0	MEDICAL	C	35000	14.27	1	0.55

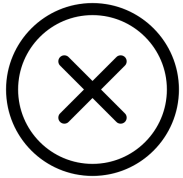
- The original dataset has 32,581 rows and 12 columns
- Our target variable is “loan_status” (1 = default; 0 = non-default)
- The 11 other columns describe a customer’s credit profile

The target variable: Loan status

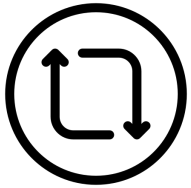


As expected, most of the data represents **“no default” customers**. We handle this imbalance during our analysis with **resampling methods to ensure robustness**.

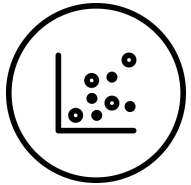
Data preprocessing and cleaning summary



Filling invalid entries



Transforming categorical variables



Removing highly correlated variables

Dataset post-processing and cleaning

Below is a table describing the columns in our final dataset

	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length	person_home_ownership_OTHER	person_home_ownership_OWN
count	3.258100e+04	32581.000000	32581.000000	32581.000000	32581.000000	32581.000000	32581.000000	32581.000000	32581.000000
mean	6.607485e+04	4.788803	9589.371106	11.011533	0.218164	0.170203	5.804211	0.003284	0.079310
std	6.198312e+04	4.087974	6322.086646	3.081605	0.413006	0.106782	4.055001	0.057214	0.270226
min	4.000000e+03	0.000000	500.000000	5.420000	0.000000	0.000000	2.000000	0.000000	0.000000
25%	3.850000e+04	2.000000	5000.000000	8.490000	0.000000	0.090000	3.000000	0.000000	0.000000
50%	5.500000e+04	4.000000	8000.000000	11.010000	0.000000	0.150000	4.000000	0.000000	0.000000
75%	7.920000e+04	7.000000	12200.000000	13.110000	0.000000	0.230000	8.000000	0.000000	0.000000
max	6.000000e+06	123.000000	35000.000000	23.220000	1.000000	0.830000	30.000000	1.000000	1.000000

The dataset we'll use in our analysis has **32,581 rows and 22 columns**. The additional columns come from assigning numerical values to categorical variables (i.e., home ownership status)

Results: Optimal model

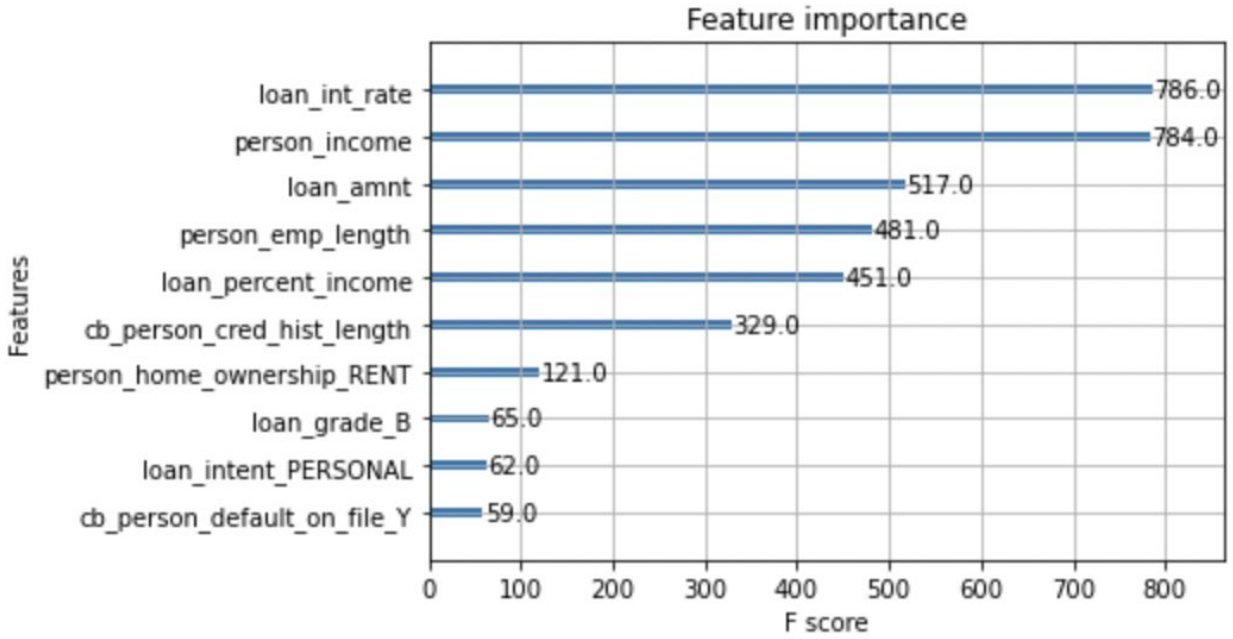
Among the different algorithms, we found our ideal model to be the tuned XGBoost

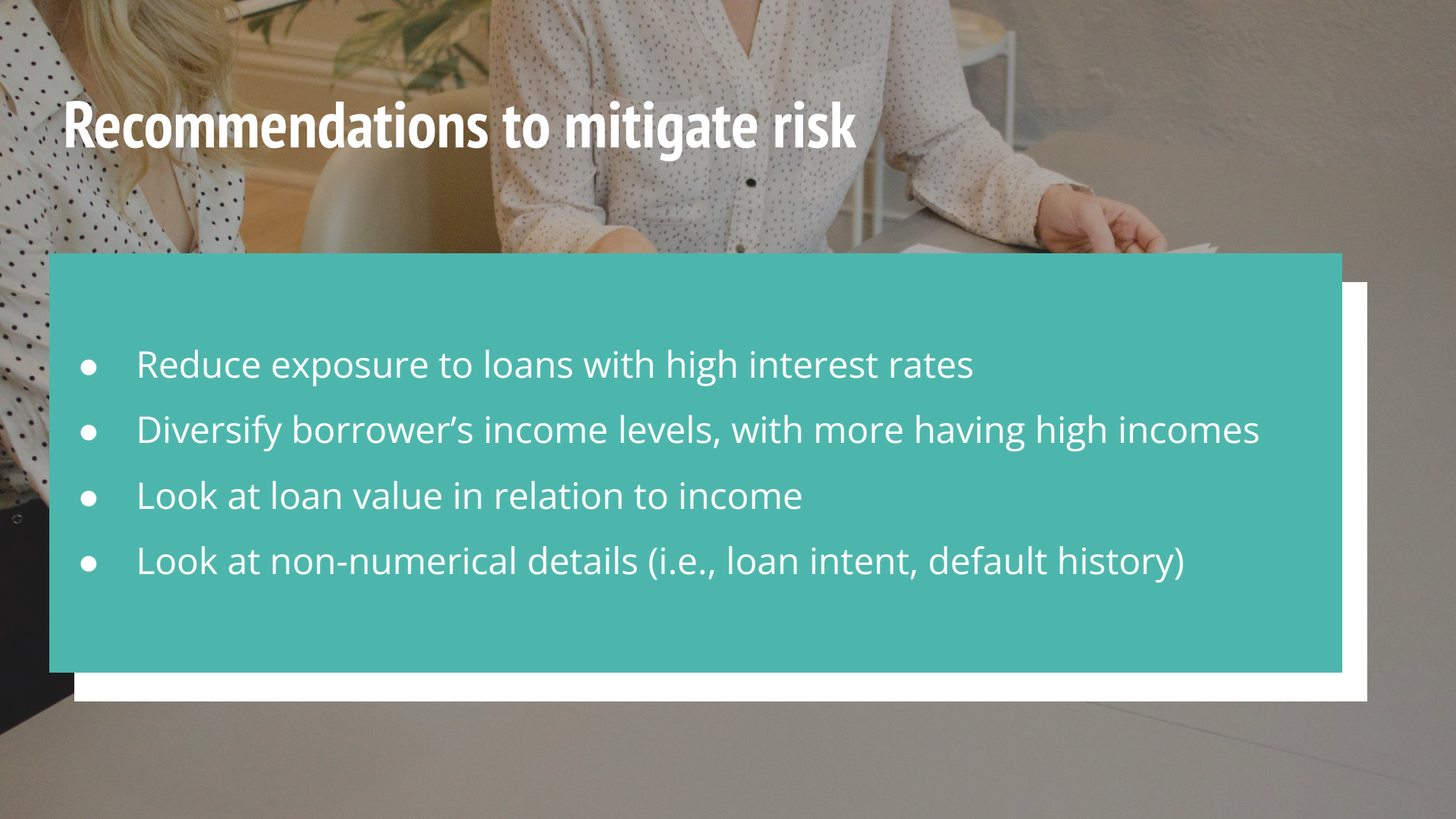
Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.72	0.71	0.75	0.72
kNN	0.83	0.81	0.86	0.83
Random Forest	0.94	0.97	0.90	0.93
XGBoost	0.94	0.96	0.92	0.94

- We tested **four** different algorithms
- Our **XGBoost** tuned to **minimize false positives** has the best performance
- This is reflected under “recall”, which is **92%** for the XGBoost model

Results: Feature importance

The barplot visualizes which features were most frequently used by our model (F-score)





Recommendations to mitigate risk

- Reduce exposure to loans with high interest rates
- Diversify borrower’s income levels, with more having high incomes
- Look at loan value in relation to income
- Look at non-numerical details (i.e., loan intent, default history)

Contact Information

Github
LinkedIn
Email