Logistic Regression & Random Forest Classification

Assignment 4
Lawrence Technological University
Department of Math and Computer Science
MCS 5623 Machine Learning
Michael Giuliani

Lawrence Tech
THEORY AND PRACTICE
1932

Preprocessing

1.) Changed “X#” column name format to use human readable names from the description row. (i.e. “X1” renamed “LIMIT_BAL”)
2.) Removed the redundant description row after renaming columns.
3.) Rename column “PAY_0” to “PAY_1” to match naming scheme of the other columns.
4.) Removed the unneeded “ID” column that sequentially numbered the rows.
5.) Checked for duplicates and dropped any found (35 rows).
6.) Checked for duplicates and found none.
7.) All data already numerically encoded.

Lawrence Tech
THEORY AND PRACTICE
1932

Data Evaluation

Example Rows

LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4
20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0
120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272
90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331
50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314
50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940

Example Rows

2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
	689	0	0	0	0	689	0	0	0	0	1
	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

Data Evaluation

Missing Values

Column Name	Null Count
LIMIT_BAL	0
SEX	0
EDUCATION	0
MARRIAGE	0
AGE	0
PAY_1	0
PAY_2	0
PAY_3	0
PAY_4	0
PAY_5	0
PAY_6	0
BILL_AMT1	0
BILL_AMT2	0
BILL_AMT3	0
BILL_AMT4	0
BILL_AMT5	0
BILL_AMT6	0
PAY_AMT1	0
PAY_AMT2	0
PAY_AMT3	0
PAY_AMT4	0
PAY_AMT5	0
PAY_AMT6	0

Data Evaluation

Basic Description Info

Description	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT
count	29965	29965	29965	29965	29965	29965	29965	29965	29965	29965	29965	29965	29965	29965
unique	81	2	7	4	56	11	11	11	11	10	10	22723	22346	22026
top	50000	2	2	2	29	0	0	0	0	0	0	0	0	0
freq	3363	18091	14019	15945	1602	14737	15730	15764	16455	16947	16286	1978	2476	2840

Duplicate Value Count

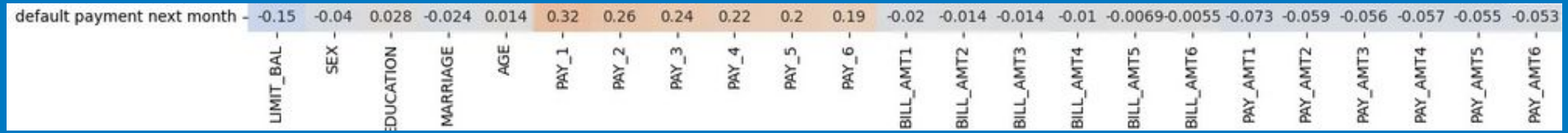
0

Class Representation

default payment next month	count
Not Defaulted	23335
Defaulted	6630

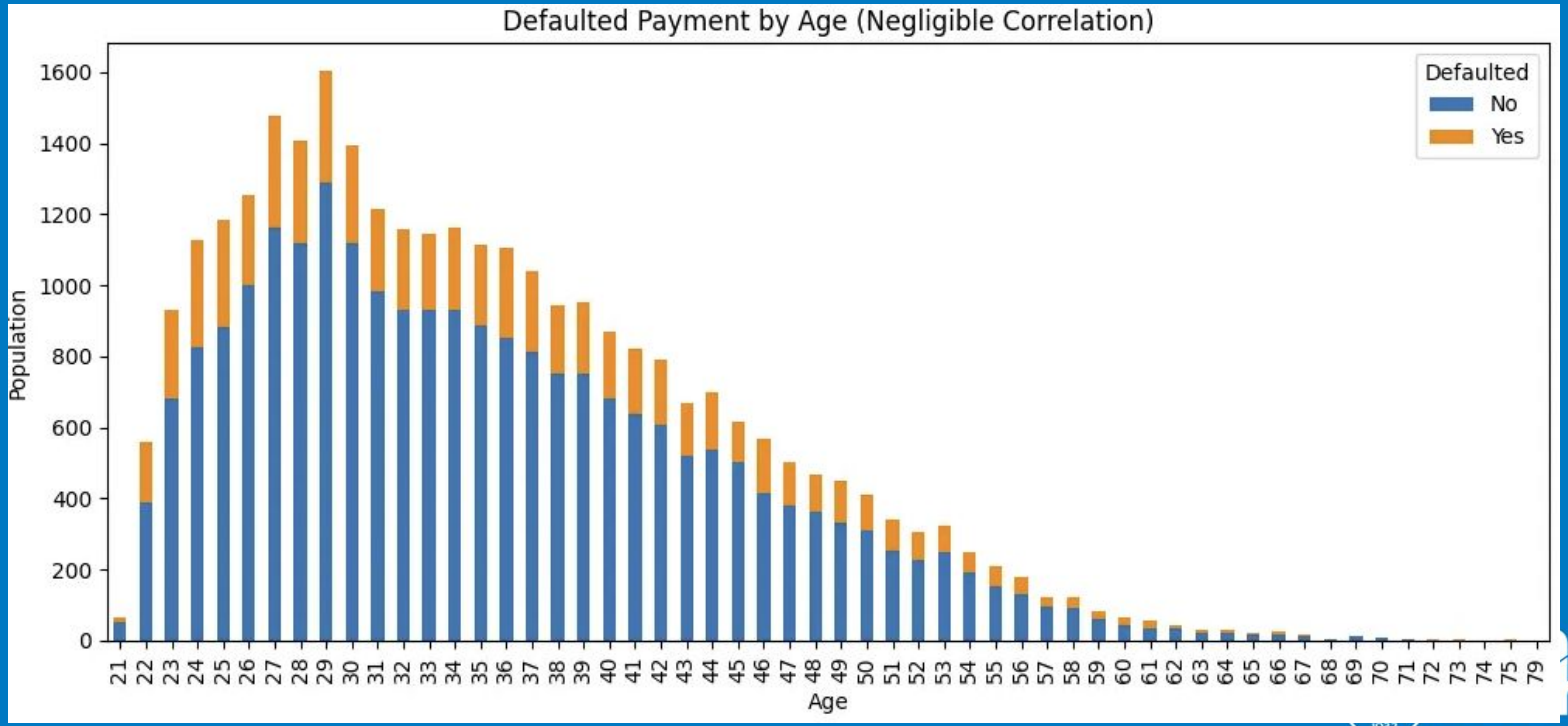
Lawrence Tech
THEORY AND PRACTICE 1932

Correlation Heatmap

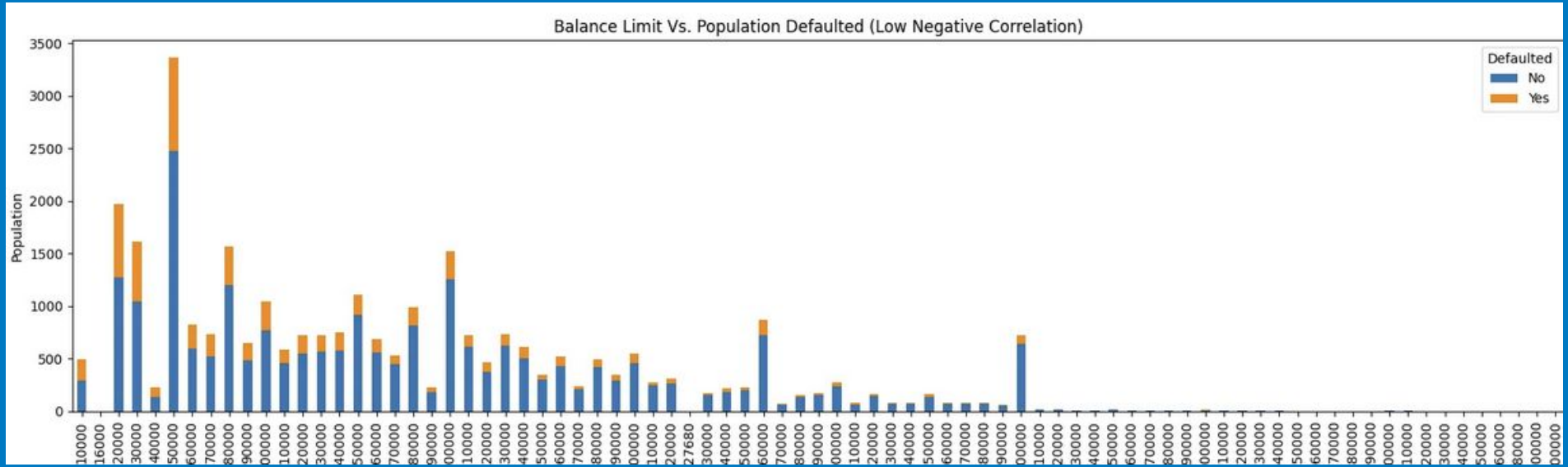


Lawrence Tech
THEORY AND PRACTICE 1932

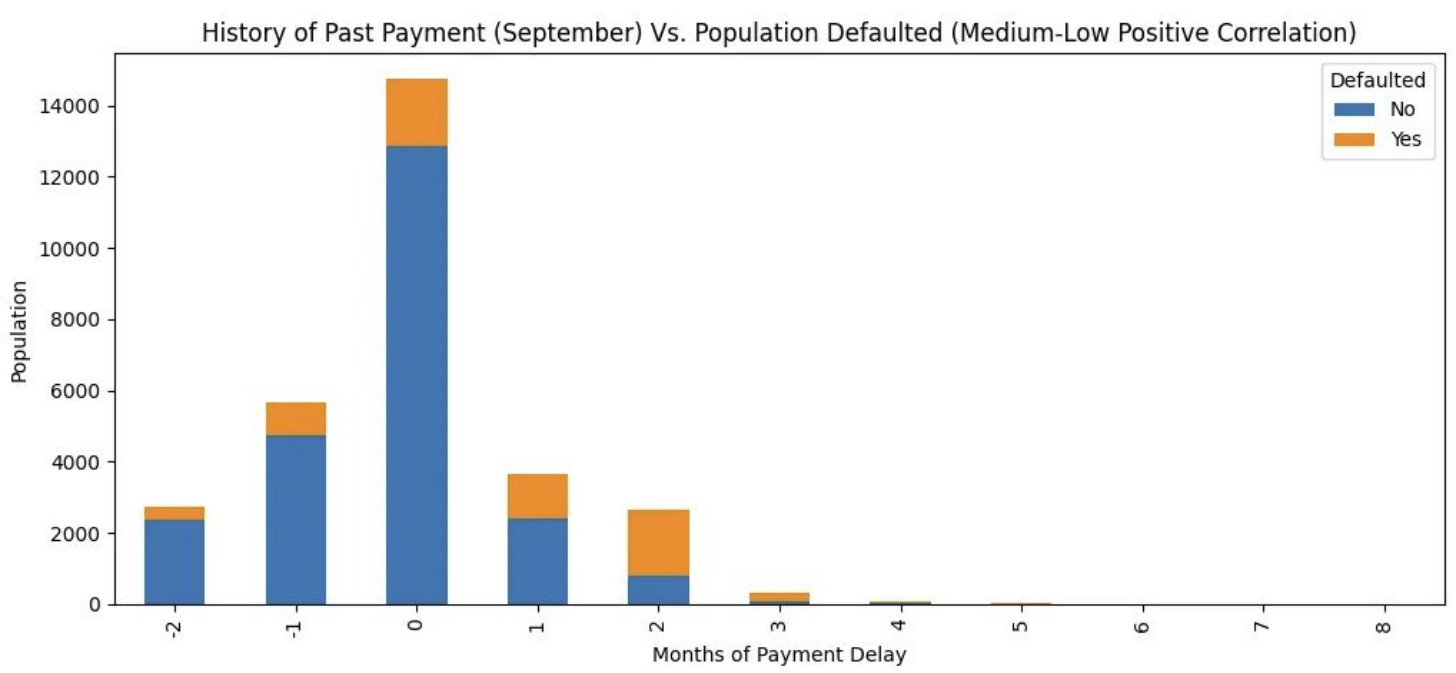
Age Distribution & Default Payments by Age (Negligible Correlation)



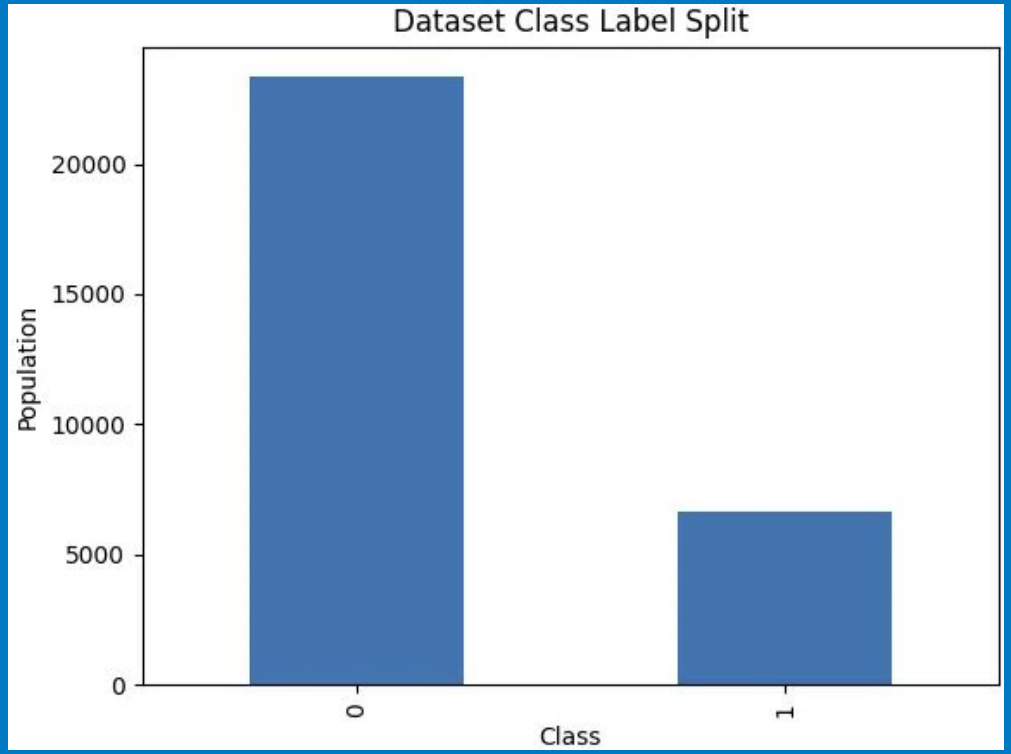
Balance Limit Vs. Default Payment Split (Small Negative Correlation)



History of Past Payment Vs. Default Payment Split (Medium-Low Positive Correlation)



Class Label Overall Split



Lawrence Tech
THEORY AND PRACTICE
1932

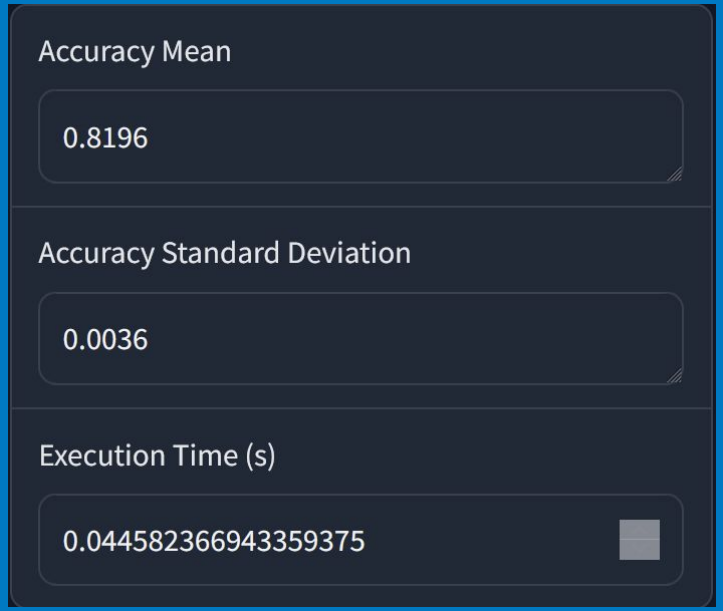
Logistic Regression

1.) Retrieve columns selected for training/testing.
2.) Scale data (mix/max or standard).
3.) Split data using stratified K-Fold to compensate for class imbalance.
    a.) Used 5 folds and data shuffling after each epoch.
4.) Create sklearn LogisticRegression object.
    a.) Changing hyperparameters from default did not improve performance in most cases.
    b.) Using the “newton-cholesky” solver improved execution time by about 0.1s while maintaining accuracy & std dev.
5.) Run cross validation on the model.
6.) Calculate mean accuracy, std ev, and execution time from the results.

Lawrence Tech
THEORY AND PRACTICE 1932

Logistic Regression

Optimal Input Data	PAY_1



Lawrence Tech
THEORY AND PRACTICE 1932

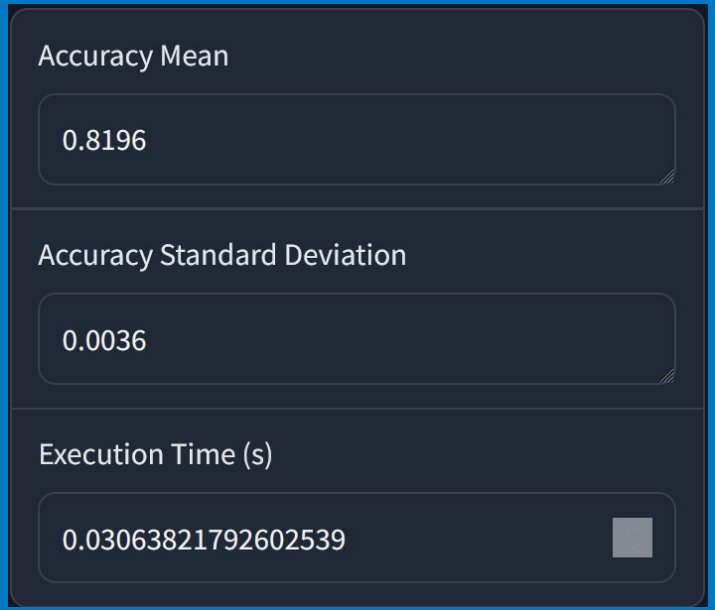
Random Forest

1.) Retrieve columns selected for training/testing.
2.) Split data using stratified K-Fold to compensate for class imbalance.
    a.) Used 5 folds and data shuffling after each epoch.
3.) Create sklearn RandomForestClassifier object.
    a.) Changing hyperparameters from default did not improve performance in most cases.
    b.) max_depth = 1 increased accuracy from 81.93% with default values to 81.96%.
    c.) n_estimators = 1 decreased execution time by about 60ms over default.
4.) Run cross validation on the model.
5.) Calculate mean accuracy, std ev, and execution time from the results.

Lawrence Tech
THEORY AND PRACTICE 1932

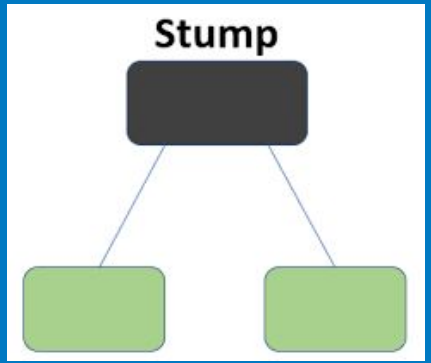
Random Forest

Optimal Input Data	PAY_1



Random Forest

- Using any columns other than “PAY_1” decreases performance.
  - “PAY_1” has the highest correlation to class labels.
- Best performance when hyperparameters create a single decision “stump”.
- Generating multiple trees decreases performance.
- Tree depth beyond 1 decreases performance.
- Useful data for classification is too simple for random forest to make sense.



Lawrence Tech
THEORY AND PRACTICE
1932

Questions?

Lawrence Tech
THEORY AND PRACTICE
1932