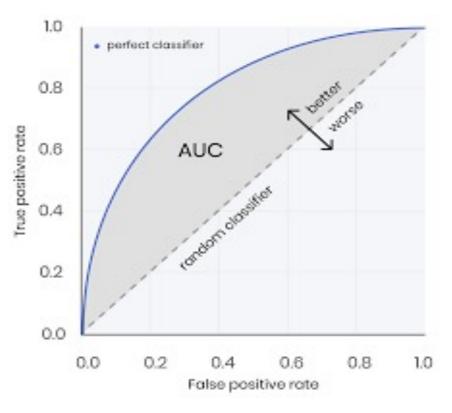
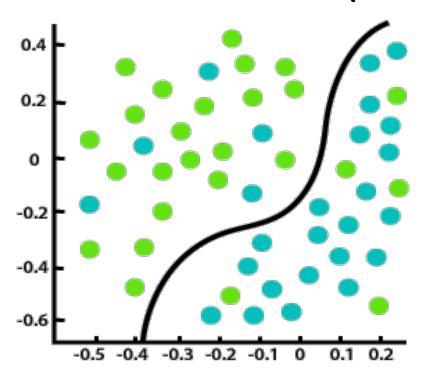
Project for STA6247.

Classification of default event(loan).





Antasiuk Vladimir

The problem: classification of non-default event (loan).

Default vs non-default events (loans).

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

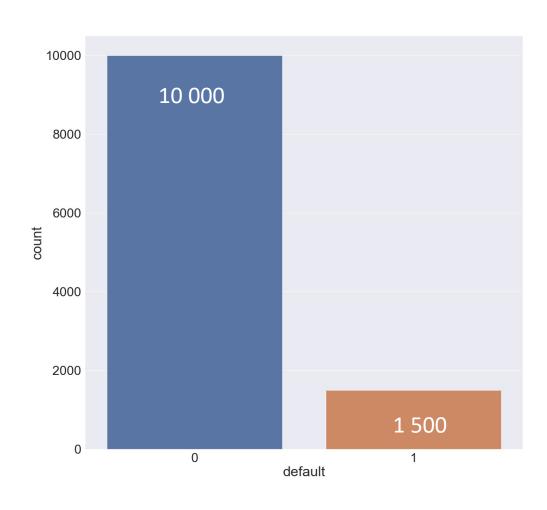
The data

Data obtained from a private company

30 encrypted features have encoded names like A1, A2 ... but we don't know what does it really mean)

Train and test sets the same size (11500 rows, 30 features) Target class: default=1, nondefault=0. 13% of defaults in each dataset

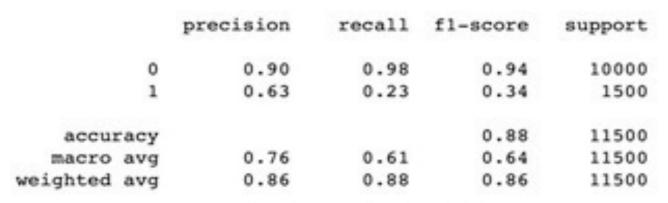
Distribution of classes in train and test datasets

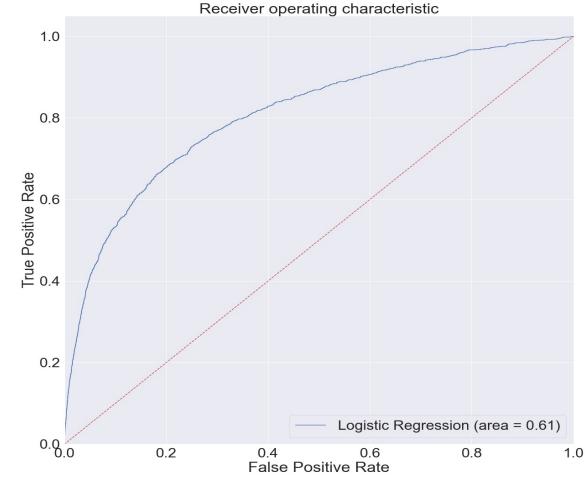


- A1	1	-0.77	-0.5	-0.4	0.52	0.51	0.53	-0.41	-0.36	-0.33	-0.5	-0.4	0.42	0.41	-0.33	0.48	-0.29	0.44	0.36	0.31	-0.012	0.018	-0.0034	0.024	0.03	-0.058	0.024	-0.016	-0.03	-0.038	0.38
-	-0.77	1	0.54	0.42	-0.52	-0.51	-0.62	0.41	0.44	0.42	0.79	0.4	-0.42	-0.41	0.4	-0.37	0.3	-0.35	-0.28	-0.25											-0.35
- A3	-0.5	0.54	1	0.67	-0.54	-0.58	-0.46	0.59			0.44		-0.47	-0.4		-0.29		-0.27	-0.24	-0.36		-0.051			-0.0046			0.043		0.046	-0.28
₩ -	-0.4		0.67	1	-0.34	-0.43	-0.39	0.4					-0.36	-0.28		-0.23		-0.21	-0.19	-0.22		-0.069							0.0088		-0.21
- A5	0.52	-0.52	-0.54	-0.34	1	0.91	0.4	-0.47	-0.32	-0.3	-0.39	-0.43	0.71	0.7	-0.2	0.25	-0.27			0.38	-0.00048					-0.048			-0.014		0.31
A6 -	0.51	-0.51	-0.58	-0.43	0.91	1		-0.49	-0.36	-0.33	-0.38	-0.44		0.66	-0.24		-0.27				0.0021		0.00064	0.042		-0.051			-0.0069		0.3
. A7	0.53	-0.62	-0.46	-0.39	0.4	0.41	1	-0.31	-0.37	-0.33	-0.49	-0.3		0.33	-0.37		-0.25				0.0013						0.041	0.0067	-0.0042	-0.0073	0.26
84 -	-0.41	0.41	0.59	0.4	-0.47	-0.49	-0.31	1	0.35	0.31	0.33	0.85	-0.37	-0.47		-0.2		-0.18	-0.16	-0.54	0.0061	-0.008					0.0061				-0.25
₽ -	-0.36				-0.32	-0.36	-0.37	0.35	1	0.9		0.28	-0.32	-0.12	0.57			-0.17	-0.19	-0.15		-0.062		-0.068			-0.0058				-0.22
A10	-0.33				-0.3	-0.33	-0.33	0.31	0.9	1			-0.29	-0.071		-0.19		-0.16	-0.18	-0.12							0.0083	0.042	0.044		-0.2
A11	-0.5	0.79			-0.39	-0.38	-0.49	0.33	0.36	0.35	1		-0.33	-0.28		-0.25		-0.24	-0.19	-0.18			-0.0003			0.047			0.046	0.00076	-0.27
A12	-0.4				-0.43	-0.44	-0.3	0.85			0.3	1	-0.34	-0.49		-0.17		-0.14	-0.14	-0.54	-0.0063	-0.016			-0.008						-0.22
A13	0.42	-0.42	-0.47	-0.36	0.71			-0.37	-0.32	-0.29	-0.33	-0.34	1	0.49	-0.22		-0.26									-0.043					0.24
A14	0.41	-0.41	-0.4	-0.28	0.7	0.66	0.33	-0.47	-0.12	-0.071	-0.28	-0.49	0.49	1	0.14		-0.14			0.46			-0.0048					0.0057			0.22
A15	-0.33	0.4	0.54	0.58	-0.2	-0.24	-0.37		0.57	0.56			-0.22	0.14	1	-0.17		-0.16	-0.16			-0.096		-0.11							-0.18
A16	0.48	-0.37	-0.29	-0.23				-0.2	-0.2	-0.19	-0.25	-0.17			-0.17	1	-0.17	0.89	0.8	0.24		-0.0086		-0.0086	-0.0031	-0.044	-0.0025				0.21
A17	-0.29	0.3	0.36	0.35	-0.27	-0.27	-0.25	0.26					-0.26	-0.14	0.38	-0.17	1	-0.16	-0.15					-0.043							-0.16
A18		-0.35	-0.27	-0.21				-0.18	-0.17	-0.16	-0.24	-0.14			-0.16	0.89	-0.16	1	0.5	0.2		-0.0053	-0.0097		-0.0022		0.00087				0.2
A19		-0.28	-0.24	-0.19	0.21			-0.16	-0.19	-0.18	-0.19	-0.14		0.15	-0.16	0.8	-0.15	0.5	1	0.18	-0.011					-0.043	-0.0063		-0.029		0.16
A20		-0.25	-0.36	-0.22	0.38	0.39		-0.54	-0.15	-0.12	-0.18	-0.54		0.46						1	0.015	-0.021			0.0082		-0.0097		-0.031		0.17
A21					-0.00048			0.0061				-0.0063						-0.013			1	0.39	0.3	0.31	0.065						-0.0021
A22				-0.069							-0.032				-0.096	-0.0086		-0.0053				1	0.7	0.88	0.64	0.24					0.016
A23	0.0034			-0.028	0.0012	0.00064					-0.0003		0.0054	-0.0048	-0.055			-0.0097	-0.019			0.7	1	0.71			-0.039				-0.00081
A24				-0.069		0.042		-0.0022	-0.068						-0.11	-0.0086	-0.043	-0.0039				0.88	0.71	1	0.63						0.018
A25			-0.0046									-0.008			0.006			-0.0022	-0.0058	0.0082	0.065	0.64		0.63	1	0.083		0.087	0.062	0.029	0.00044
A26	-0.058				-0.048						0.047		-0.043			-0.044			-0.043				0.52		0.083	1	-0.077			0.66	-0.034
A27	0.024	-0.027	-0.023	-0.05	0.031	0.04	0.041	0.0061	-0.0058	0.0083	-0.0003	-0.017	0.027	0.04	-0.038	-0.0025	-0.011	0.00087	-0.0063	-0.0097	0.13	0.3	-0.039	0.33	0.077	-0.077	1	0.48	0.34	-0.071	0.0044

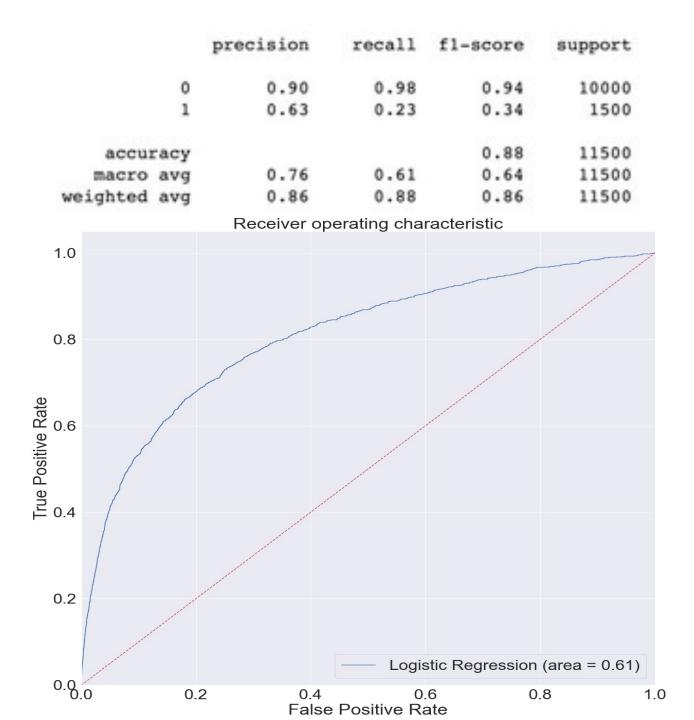
Finding and excluding highly correlated features

Logistic Regression – 10 features (train set)

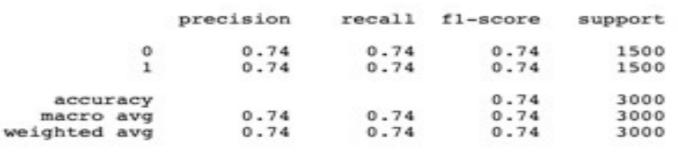


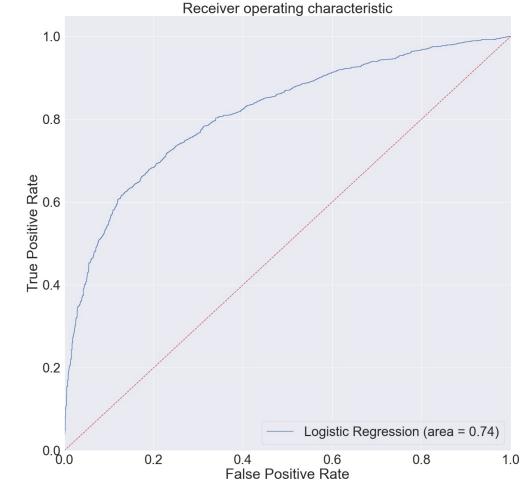


Logistic Regression – 10 features (test set)

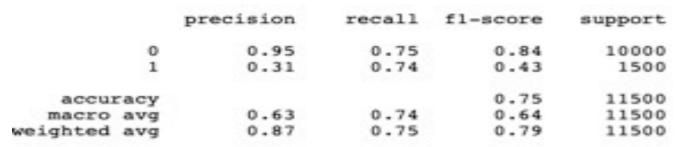


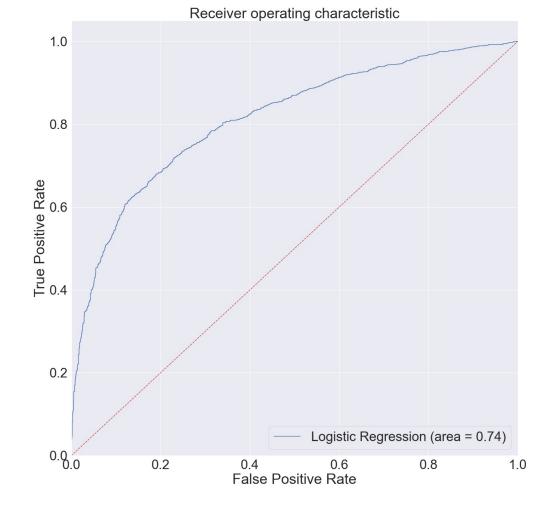
Logistic
Regression
(undersampling,
train set).
7 variables





Logistic
Regression
(undersampling, test set).
7 variables.

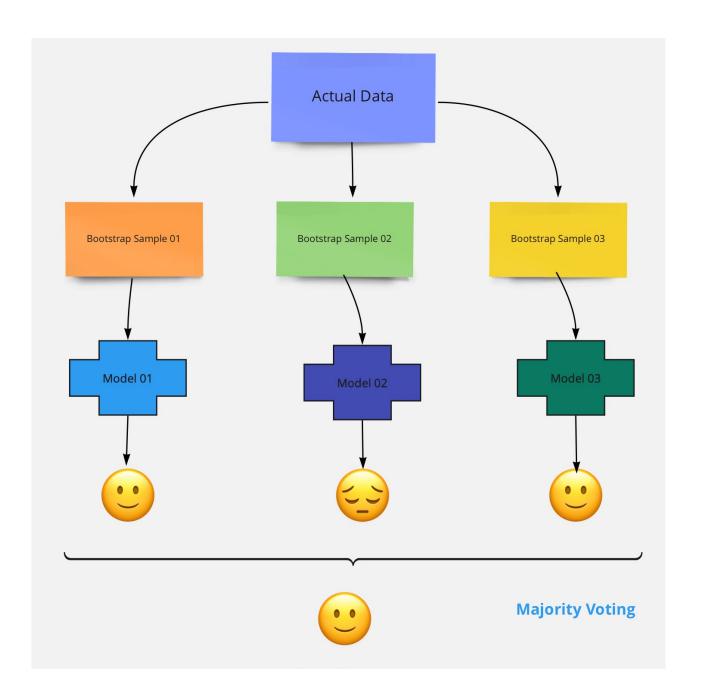




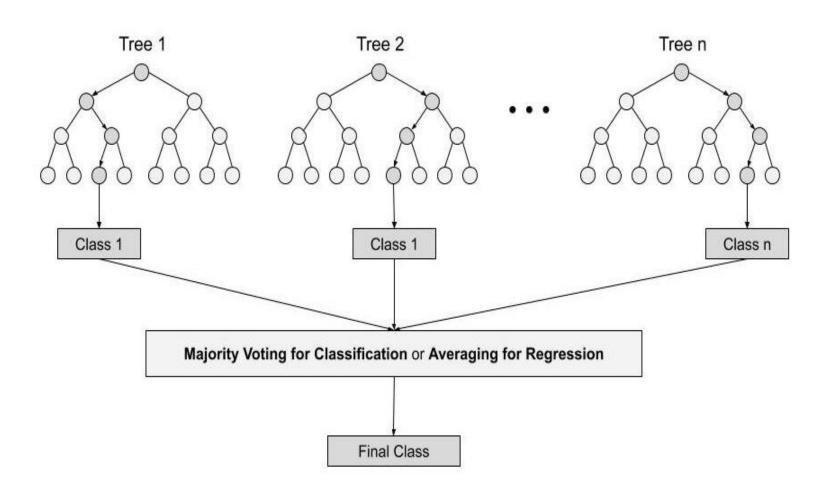
Random Forest Classifier

- Step 1: In Random forest n number of random records are taken from the data set having k number of records.
- <u>Step 2</u>: Individual decision trees are constructed for each sample.
- Step 3: Each decision tree will generate an output.
- <u>Step 4</u>: Final output is considered based on *Majority Voting or Averaging* for Classification and regression respectively.

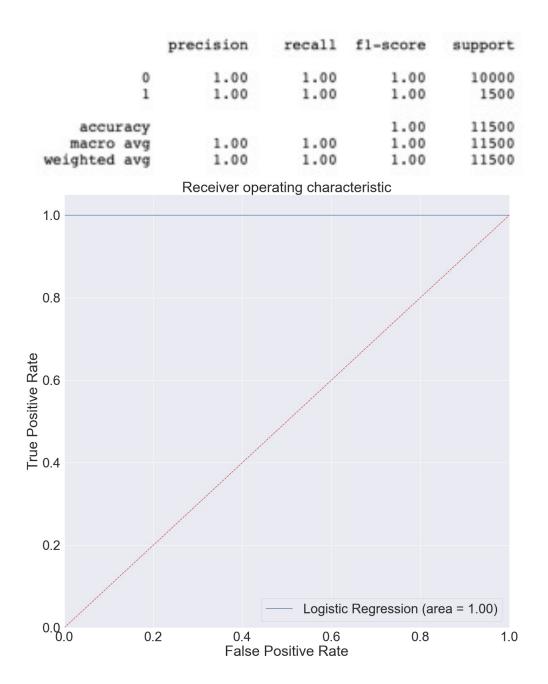
Bagging Ensemble Method



Random Forest Classifier



Random Forest Classifier (same confusion matrix and AUC on train and test sets)



Conclusion

- We don't need all 30 variables to build sufficient logistic regression model
- Undersampling/oversampling can help improve model precision (in our case AUC improved from 0.61 to 0.74 with lighter model)
- Sophisticated algorithms can improve model quality (Random forest classifier gave us perfect AUC)