JANUARY 2021

MODELLING CHALLENGE: FRAUD DETECTION

ALLISON BLACK

IE MBD
FINANCIAL, FRAUD, AND RISK ANALYTICS
PROF. MANOEL FERNANDO ALONSO GADI

MODELLING CHALLENGE LOGBOOK

Attempt	Date	Gini	KS2	Notes
1	11/01	0.11813	0.12548	I ran Professor Manoel's starting code as is (no changes)
2	21/01	0.11813	0.12548	I produced a heatmap of the DataFrame, so that I can easily see a distinction between the different types of variables and to see if there is skewedness in outputs of the variables (no changes to code or variables)
3	22/01	0.11813	0.12548	I produced a heatmap to show correlation between the variables. From session 5 class notes: "Correlation is *normal* on the balance sheet. However, we do feature selection on certain areas. We don't want one variable dominating – no more than 20% of feature importance." (no change to code or variables)
4	22/01	0.10932	0.11808	Researching how to deal with the imbalanced dataset: *from imblearn import RandomUnderSampler*
5	22/01	0.29291	0.24744	I changed the *in_model* to ['ib_var_1','ib_var_2','ib_var_4','ib_var_5','ib_var_8','ib_var_18']
6	24/01	0.28748	0.24834	I changed the *in_model* to ['ib_var_1','ib_var_2','ib_var_4','ib_var_5','ib_var_6', 'icn_var_22','ico_var_25','if_var_65']
7	24/01	0.11813	0.12548	Mistakenly ran original code
8	24/01	0.28836	0.24607	I implemented RandomForestClassifier instead of logistic regression (no change to variables). I chose RandomForest because it most of the previous machine learning projects, it seems to be the algorithm with the best results.
9	24/01	0.28905	0.24682	Along with *RandomForestClassifier*, I also ran the original *LogisticRegression* (no change to variables)
10	24/01	0.34921	0.27013	Error in dfo: it's showing 83 columns. I will remove my current input data and re-upload oot0.csv and dev.csv
11	24/01	0.29363	0.25951	I changed the *in_model* to ['ib_var_1','ib_var_4','ib_var_5','ib_var_6','icn_var_22', 'ico_var_25','if_var_65'] (removed ib_var2): lower score!
12	24/01	0.34257	0.26331	I changed the *in_model* to the original, then back to what I had in step 6. Higher score!
13	24/01	0.34378	0.26455	Added back ib_var2 Removed ib_var_3, and added it back in
14	24/01	0.33247	0.26070	['ib_var_1','ib_var_2','ib_var_4','ib_var_5','ib_var_6','ib_var_7', 'ib_var_8', 'ib_var_9','icn_var_22','ico_var_25','if_var_65']
15	24/01	0.33355	0.27106	Again, trying to deal with imbalanced dataset: *from imblearn.over_sampling import RandomOverSampler*

MODELLING CHALLENGE LOGBOOK

16	24/01	0.33925	0.26646	"ValueError: Found input variables with inconsistent numbers of samples: [864, 1552]" I printed the following shapes to investigate and change variables: X shape: (864, 8) y shape: (864,) pred_dev shape: (1552,)
17	24/01	0.34476	0.26555	I selected all variables for *in_model*. Lower score
18	24/01	0.50794	0.37707	I kept all variables selected for *in_model* but removed *LogisticRegression* and kept *RandomForestClassifier*
19	24/01	0.51559	0.38386	Fixed error in step 4: *gini_score* Final Submission!

	username	gini	ks2	grade	received_at
0	a.black	0.51559	0.38386	9.09600	2021-01-24 21:37:19
1	a.black	0.50794	0.37707	8.93500	2021-01-24 21:22:42
2	a.black	0.50794	0.37707	8.93500	2021-01-24 21:06:34
3	a.black	0.34476	0.26555	6.29300	2021-01-24 19:46:26
4	a.black	0.34231	0.26369	6.24800	2021-01-24 19:39:46
5	a.black	0.33925	0.26646	6.31400	2021-01-24 19:39:12
6	a.black	0.33355	0.27106	6.42300	2021-01-24 19:15:10
7	a.black	0.29011	0.24834	5.88500	2021-01-24 19:13:42
8	a.black	0.34838	0.26582	6.29900	2021-01-24 19:10:18
9	a.black	0.33247	0.26070	6.17800	2021-01-24 19:05:23
10	a.black	0.34378	0.26455	6.26900	2021-01-24 17:23:25
11	a.black	0.34257	0.26331	6.23900	2021-01-24 17:22:09
12	a.black	0.33954	0.27064	6.41300	2021-01-24 17:20:29
13	a.black	0.29363	0.25951	6.15000	2021-01-24 17:18:26
14	a.black	0.34921	0.27013	6.40100	2021-01-24 10:53:05
15	a.black	0.13855	0.10549	2.50000	2021-01-24 10:51:58
16	a.black	0.28905	0.24682	5.84900	2021-01-24 10:50:17
17	a.black	0.29293	0.24531	5.81300	2021-01-24 10:49:37
18	a.black	0.28836	0.24607	5.83100	2021-01-24 10:40:59
19	a.black	0.11813	0.12548	2.97400	2021-01-24 10:40:19
20	a.black	0.28748	0.24834	5.88500	2021-01-24 10:33:40
21	a.black	0.29291	0.24744	5.86400	2021-01-22 17:23:51
22	a.black	0.10932	0.11808	2.79800	2021-01-22 17:17:37
23	a.black	0.11813	0.12548	2.97400	2021-01-22 16:31:44
24	a.black	0.11813	0.12548	2.97400	2021-01-21 16:03:29
25	a.black	0.11813	0.12548	2.97400	2021-01-11 10:58:10

RESOURCES

Fraud detection ideas:
https://www.youtube.com/watch?v=8PU4N8EeQPc

Manoel Fernando Alonso Gadi, Alair Pereira do Lago and Jorn Mehnen (February 1st 2010). Data Mining with Skewed Data, New Advances in Machine Learning, Yagang Zhang, IntechOpen, DOI: 10.5772/9382. Available from: https://www.intechopen.com/books/new-advances-in-machine-learning/data-mining-with-skewed-data

Fernando Alonso Gadi, M., Wang, X., & Pereira do Lago, A. (n.d.). Credit Card Fraud Detection with Artificial Immune System (Tech.). Retrieved January 25, 2021, from https://www.ime.usp.br/~mgadi/paper_CreditCardFraudDetectionWithArtificialImmuneSystem_ICARIS08.pdf

Fernando Alonso Gadi, M., Wang, X., & Pereira do Lago, A. (n.d.). Comparison with Parametric Optimization in Credit Card Fraud Detection (Tech.). Retrieved from https://www.ime.usp.br/~mgadi/paper_ComparisonwithParametricOptimizationinCreditCardFraudDetection_ICMLA08.pdf

cde48001122row14_col49,#T_a2a3c728_5e79_11eb_b80f_acde48001122row14_col51,#T_a2a3c728_5e79_11eb_b80f_acde4
8001122row14_col64,#T_a2a3c728_5e79_11eb_b80f_acde48001122row14_col70,#T_a2a3c728_5e79_11eb_b80f_acde48001
122row14_col74,#T_a2a3c728_5e79_11eb_b80f_acde48001122row14_col82,#T_a2a3c728_5e79_11eb_b80f_acde48001122r
ow15_col2,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col3,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_c
ol4,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col7,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col8,#T
_a2a3c728_5e79_11eb_b80f_acde48001122row15_col9,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col10,#T_a2a3
c728_5e79_11eb_b80f_acde48001122row15_col11,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col12,#T_a2a3c728
_5e79_11eb_b80f_acde48001122row15_col13,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col14,#T_a2a3c728_5e7
9_11eb_b80f_acde48001122row15_col18,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col23,#T_a2a3c728_5e79_11
eb_b80f_acde48001122row15_col24,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col33,#T_a2a3c728_5e79_11eb_b
80f_acde48001122row15_col70,#T_a2a3c728_5e79_11eb_b80f_acde48001122row15_col71,#T_a2a3c728_5e79_11eb_b80f_
acde48001122row15_col74,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col1,#T_a2a3c728_5e79_11eb_b80f_acde4
8001122row16_col2,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col3,#T_a2a3c728_5e79_11eb_b80f_acde480112
2row16_col4,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col5,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16
_col7,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col9,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col10
,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col12,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col13,#T_
a2a3c728_5e79_11eb_b80f_acde48001122row16_col14,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col21,#T_a2a3
c728_5e79_11eb_b80f_acde48001122row16_col23,#T_a2a3c728_5e79_11eb_b80f_acde48001122row16_col33,#T_a2a3c728