MICRO CREDIT PROJECT

SUBMITTED BY:

DEEPRO SENGUPTA

BACKGROUND

- Client is an Indonesia based telecommunications firm
- The firm looking to extend micro credit loans on talk-time in form of IDR 5 & 10 for which an amount of IDR 7 & 12 respectively is to be paid back.
- In order to do so, the client has partnered with a Microfinance Institution (MFI)

THE PROBLEM STATEMENT

- Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan.

THE DATA

- Data source: Provided by Client
- Contains 37 columns and 2,09,593 rows
- Target Variable name: ‘label’
- The dataset is imbalanced as label ‘0’ is just 12.5% while label ‘1’ is 87.5% of the rows in the dataset.

THE DATA: SUMMARY STATISTICS

	label	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma
count	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.00000	209593.000000	209593.000000
mean	0.875177	8112.343445	5381.402289	6082.515068	2692.581910	3483.406534	3755.84780	3712.202921	2064.452797
std	0.330519	75696.082531	9220.623400	10918.812767	4308.586781	5770.461279	53905.89223	53374.833430	2370.786034
min	0.000000	-48.000000	-93.012667	-93.012667	-23737.140000	-24720.580000	-29.00000	-29.000000	0.000000
25%	1.000000	246.000000	42.440000	42.692000	280.420000	300.260000	1.00000	0.000000	770.000000
50%	1.000000	527.000000	1469.175667	1500.000000	1083.570000	1334.000000	3.00000	0.000000	1539.000000
75%	1.000000	982.000000	7244.000000	7802.790000	3356.940000	4201.790000	7.00000	0.000000	2309.000000
max	1.000000	999860.755200	265926.000000	320630.000000	198926.110000	200148.110000	998650.37770	999171.809400	55000.000000

THE DATA: SUMMARY STATISTICS (CONTD)

cnt_ma_rech30	fr_ma_rech30	sumamnt_ma_rech30	medianamnt_ma_rech30	medianmarechprebal30	cnt_ma_rech90	fr_ma_rech90	sumamnt_ma_rech90
209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.00000	209593.000000	209593.000000
3.978057	3737.355121	7704.501157	1812.817952	3851.927942	6.31543	7.716780	12396.218352
4.256090	53643.625172	10139.621714	2070.864620	54006.374433	7.19347	12.590251	16857.793882
0.000000	0.000000	0.000000	0.000000	-200.000000	0.00000	0.000000	0.000000
1.000000	0.000000	1540.000000	770.000000	11.000000	2.00000	0.000000	2317.000000
3.000000	2.000000	4628.000000	1539.000000	33.900000	4.00000	2.000000	7226.000000
5.000000	6.000000	10010.000000	1924.000000	83.000000	8.00000	8.000000	16000.000000
203.000000	999606.368100	810096.000000	55000.000000	999479.419300	336.00000	88.000000	953036.000000

THE DATA: SUMMARY STATISTICS (CONTD)

medianamnt_ma_rech90	medianmarechprebal90	cnt_da_rech30	fr_da_rech30	cnt_da_rech90	fr_da_rech90	cnt_loans30	amnt_loans30	maxamnt_loans30
209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000
1864.595821	92.025541	262.578110	3749.494447	0.041495	0.045712	2.758981	17.952021	274.658747
2081.680664	369.215658	4183.897978	53885.414979	0.397556	0.951386	2.554502	17.379741	4245.264648
0.000000	-200.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
773.000000	14.600000	0.000000	0.000000	0.000000	0.000000	1.000000	6.000000	6.000000
1539.000000	36.000000	0.000000	0.000000	0.000000	0.000000	2.000000	12.000000	6.000000
1924.000000	79.310000	0.000000	0.000000	0.000000	0.000000	4.000000	24.000000	6.000000
55000.000000	41456.500000	99914.441420	999809.240100	38.000000	64.000000	50.000000	306.000000	99864.560860

THE DATA: SUMMARY STATISTICS (CONTD)

medianamnt_loans30	cnt_loans90	amnt_loans90	maxamnt_loans90	medianamnt_loans90	payback30	payback90
209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000	209593.000000
0.054029	18.520919	23.645398	6.703134	0.046077	3.398826	4.321485
0.218039	224.797423	26.469861	2.103864	0.200692	8.813729	10.308108
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	1.000000	6.000000	6.000000	0.000000	0.000000	0.000000
0.000000	2.000000	12.000000	6.000000	0.000000	0.000000	1.666667
0.000000	5.000000	30.000000	6.000000	0.000000	3.750000	4.500000
3.000000	4997.517944	438.000000	12.000000	3.000000	171.500000	171.500000

THE DATA: SUMMARY STATISTICS KEY TAKEAWAYS

- The mean is larger then the 50th percentile and there is a huge difference between the 75th percentile and max values for most of the columns. This indicates the presence of outliers.
- Minimum values of the columns are either zero or negative. The negative values are not logical for this dataset as most columns indicate values in units of days or monetary amounts which cannot be negative.

DATA PRE-PROCESSING STEPS

- Step 1: Deleting unwanted columns. The columns dropped from the dataset are ‘slno’, ‘msisdn’, ’pdate’ and ‘pcircle’.
- Step 2: Outlier Removal using the Inter-quartile Range (IQR) method
- Step 3: Data set balancing using oversampling techniques

DATA MODELLING & EVALUATION

- Algorithms used: Random Forest Classifier, Support Vector Classifier and K-Nearest Neighbors.
- Evaluation Metrics: F1-score
- Final Model Selection Criterion: The model with the least difference between the actual f1-score and the cross-validation f1-score.
- Final Model: Random Forest Classifier

DATA MODELLING & EVALUATION (CONTD.)

- F1- score for Random Forest before Hyper-parameter tuning: 94%
- F1- score for Random Forest after Hyper-parameter tuning: 94.86%

CONCLUSION

- The model is able to calculate the probability of the borrower paying back with decent accuracy.
- The probability of paying back ranges from 2% to 98%

LIMITATIONS & WAY FORWARD

- Only three classification were used out of which one was selected for final use and deployment.
- More algorithms should be tried for better results.
- The performance of the model in the real world needs to be check and the model should be adjusted as needed.