2018

Predictive Modelling for Default Detection

EDA * FEATURE ENGINEERING * HYPER-PARAMETER TUNING * MODELLING

DIANE PARK // MENTOR: SRDJAN SANTIC

Intermediate Data Science Course || Springboard

Introduction

- **Aim: extract useful and meaningful information from data**
- **Goal: build a predictive model for unseen data**
- **Primary data: applications of the loan to a loan agency**
- **Supportive data: applicants' historical transaction data**

  from agency itself, other credit card agency and others
- **Importance of historical data:**

  agency to learn the applicant's behaviour and

  to predict their repayment capability in the future

  applicants to approve their credible attitude

Analysis plan

- **Explanatory data analysis (EDA):**

  Missing data management

  Abnormality detection

  Correlation of each pair of variables

  Aggregate historical data by ID

- **Modelling:**

  Hyper-parameter tuning using cross-validation

  Feature engineering using several different algorithms

  Logistic regression and LGBM

  Scoring using area under ROC curve (AUC)

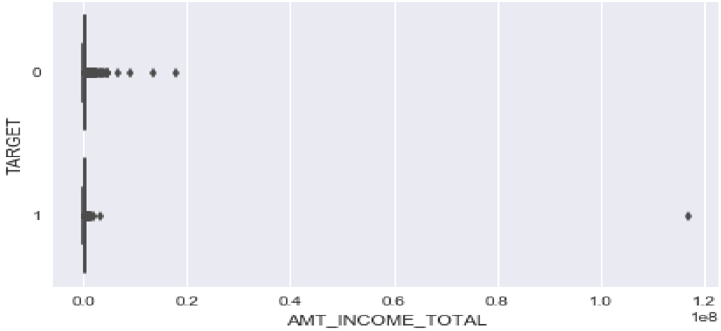
Exploratory data analysis

Dataset	Num. of rows	Num. of columns	Max. NA %	NA indicator	Outlier	Encode	Aggregate
Application	307511	122	69.9	V	V	V	
CC balance	3840312	23	20.0				V
Bureau	1716428	17	71.5				V
Bureau balance	27299925	3	11.4			V	V
Instalments	13605401	8	0.0				V
POS balance	10001358	8	0.3				V

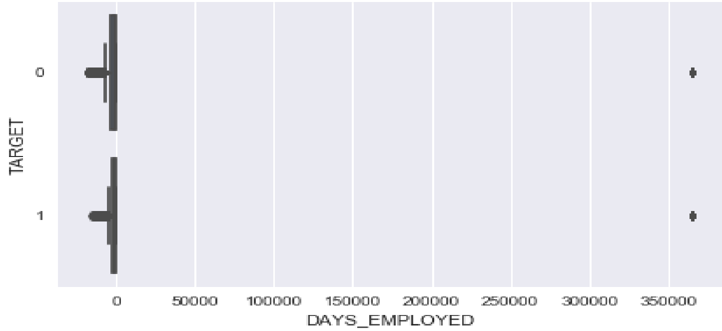
- **Default percent: 8.1% with TARGET = 1**
- **Male applicants: 34%**
- **Homeowners: 69%**

Outliers

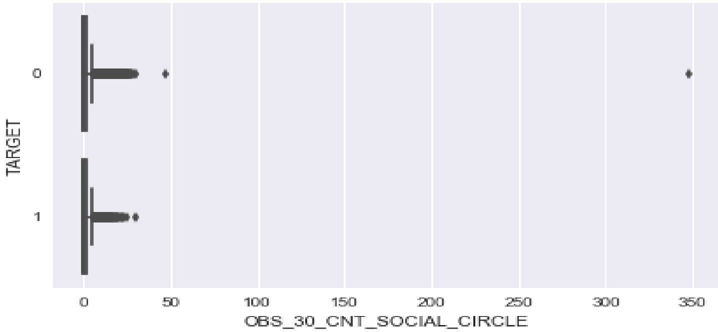
Income

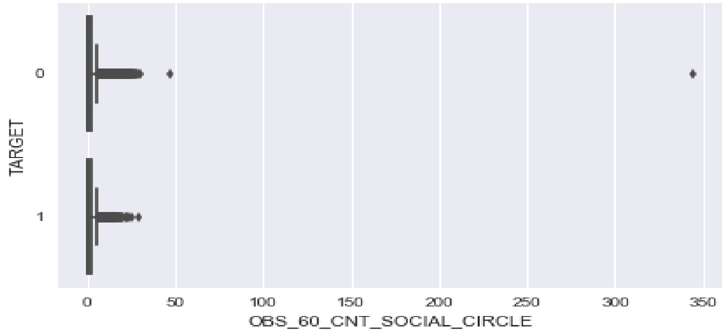


Duration of employment



Social circles in 30 days and 60 days

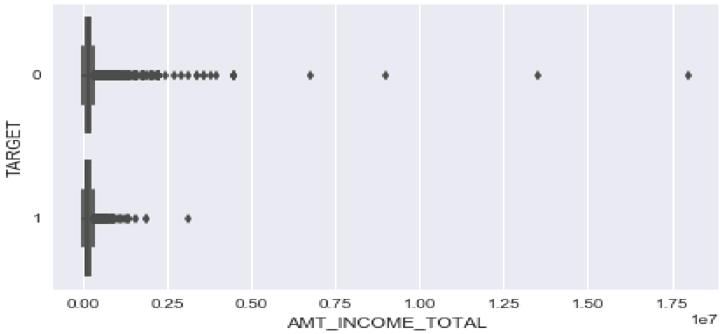




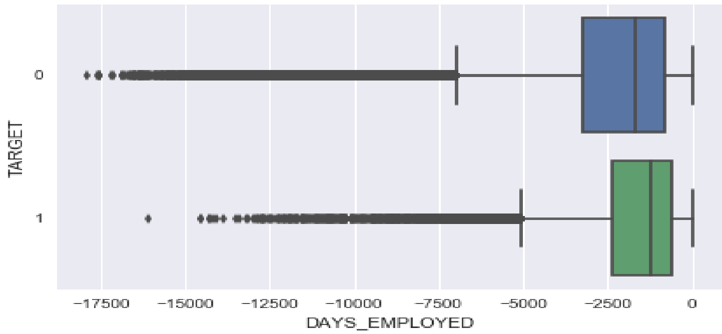
- **Income of 1.2e8, 1000 years of employment and 360 social circles**

Implementation of outlier

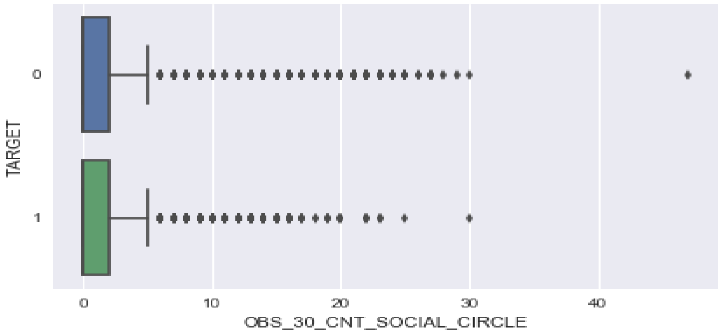
Income

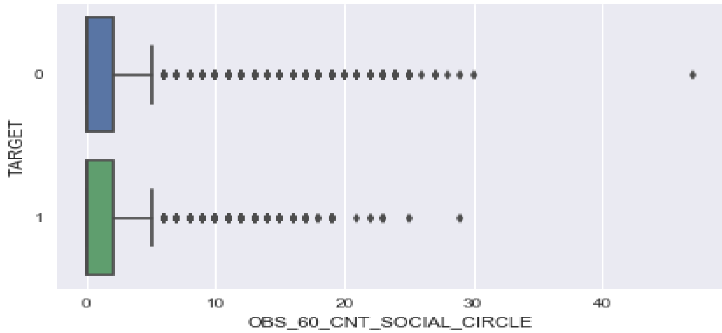


Duration of employment



Social circles in 30 days and 60 days





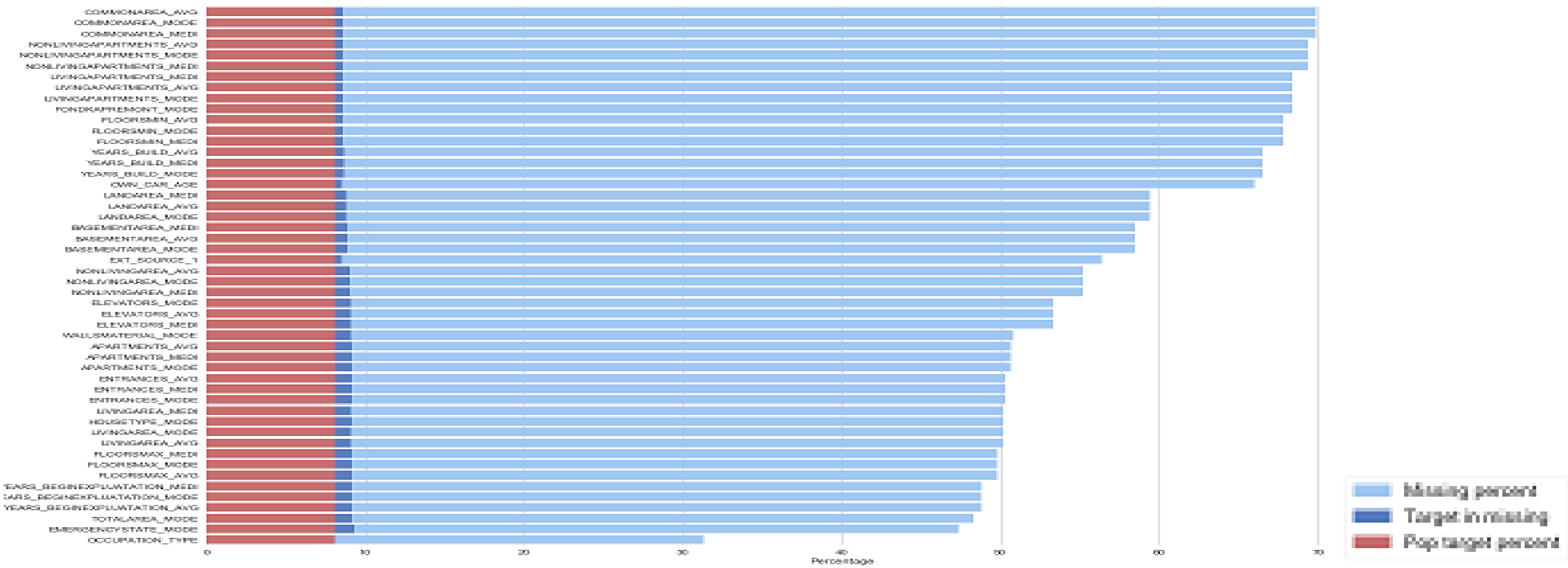
- **Abnormal values are replaced with NA**

Missing data

	variable	na_percent	na_target_percent	no_na_target_percent
61	COMMONAREA_MODE	69.9	8.6	6.9
75	COMMONAREA_MEDI	69.9	8.6	6.9
47	COMMONAREA_AVG	69.9	8.6	6.9
83	NONLIVINGAPARTMENTS_MEDI	69.4	8.6	6.9
55	NONLIVINGAPARTMENTS_AVG	69.4	8.6	6.9
69	NONLIVINGAPARTMENTS_MODE	69.4	8.6	6.9
53	LIVINGAPARTMENTS_AVG	68.4	8.6	6.9
67	LIVINGAPARTMENTS_MODE	68.4	8.6	6.9
81	LIVINGAPARTMENTS_MEDI	68.4	8.6	6.9
85	FONDKAPREMONT_MODE	68.4	8.6	6.9

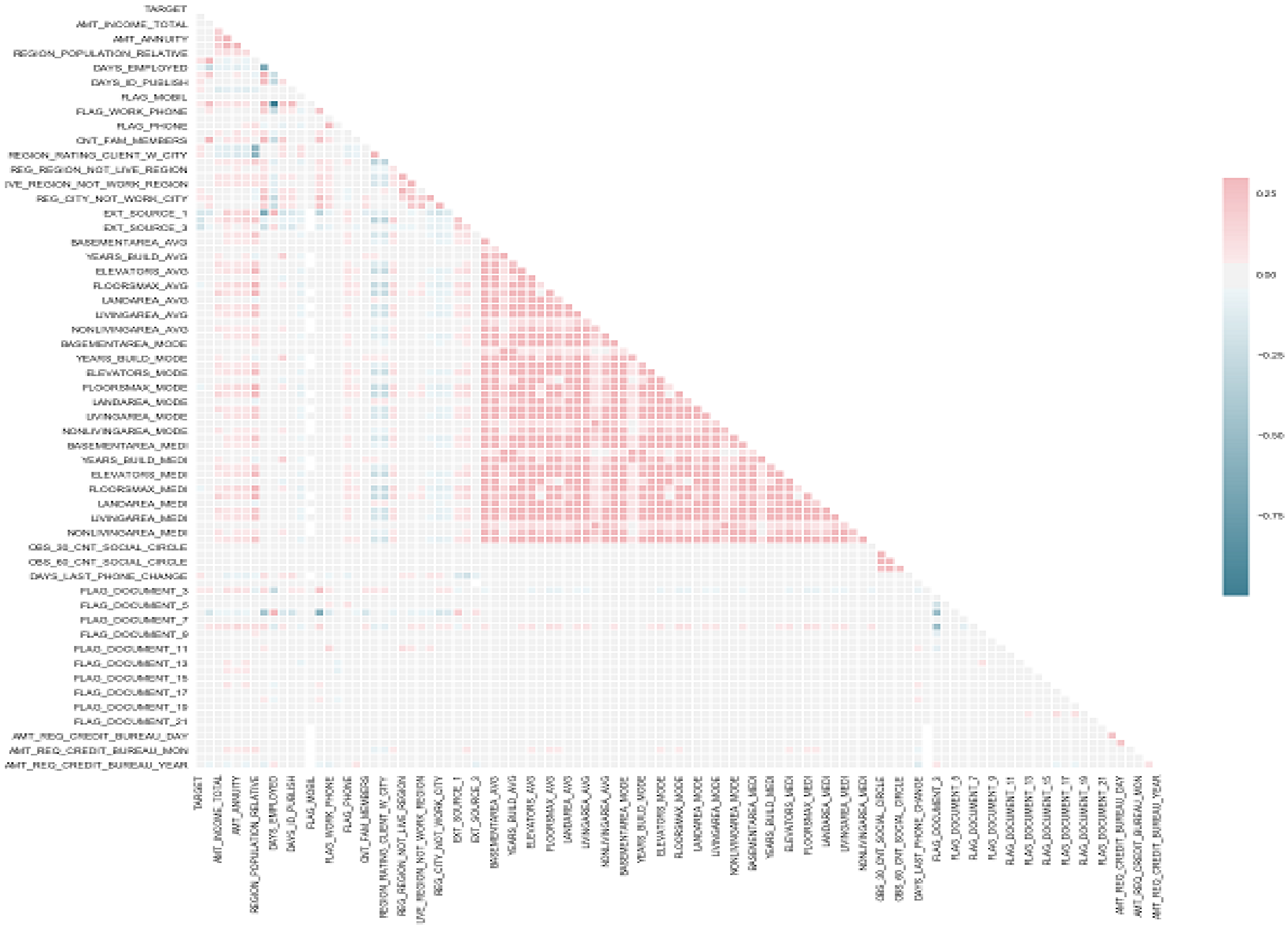
- **Missing values in 65 variables as high as 70%**
- **na_target_percent: TARGET=1 percentage in missing group**
- **no_na_target_percent: TARGET=1 percentage in non-missing group**

Difference of TARGET in groups



- **Light-blue: missing percentage**
- **Dark-blue: TARGET=1 percent in missing group**
- **Red colour: TARGET=1 percent in non-missing group**
- **Chi-square tests: confirm significant differences for 63 variables**
- **Implement indicator variables**

Correlation

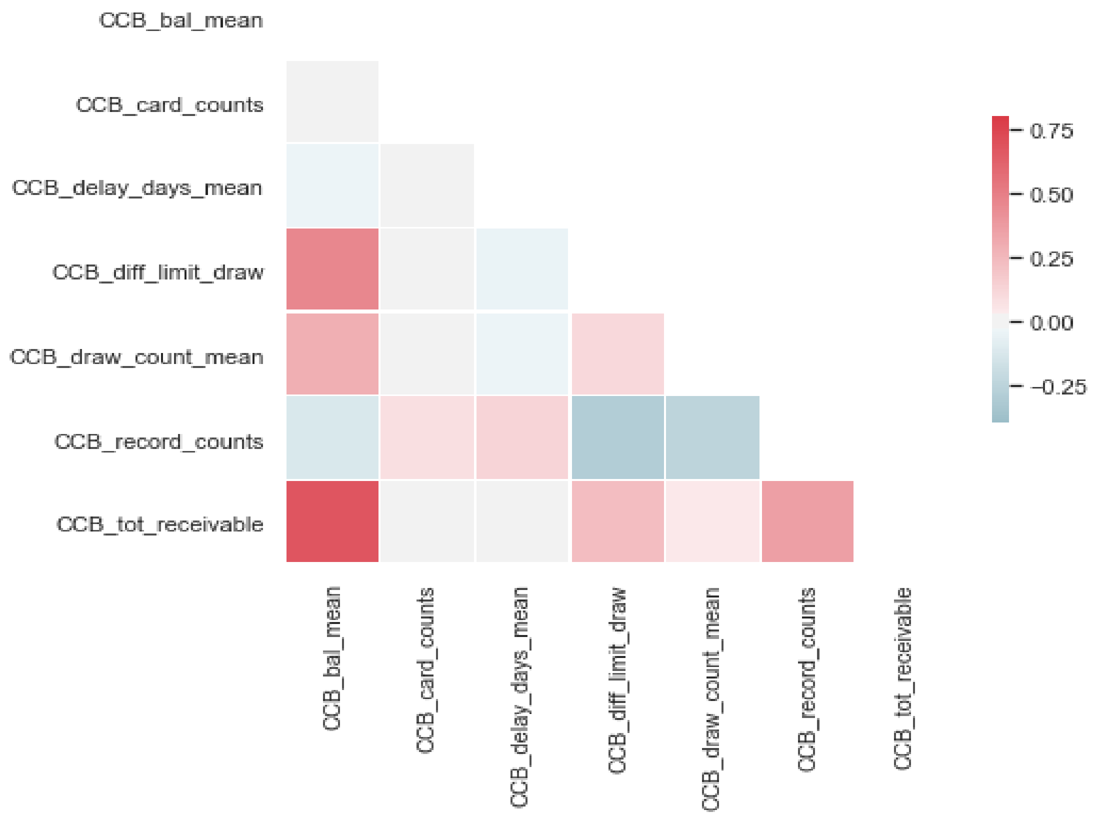


Credit card balance

- **Univariate distributions in histograms**

- **Bivariate distributions in scatter plots**

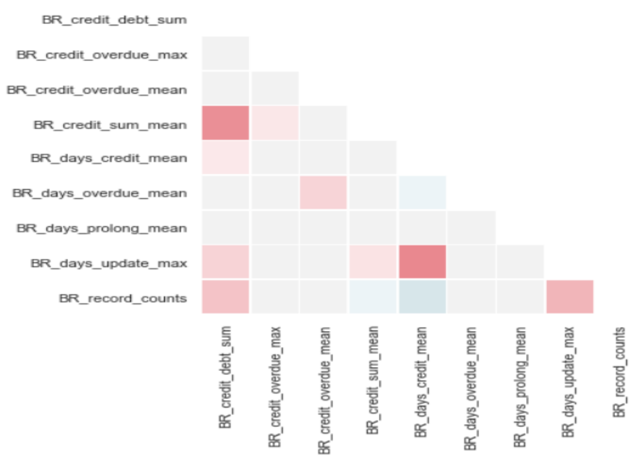
Correlation: credit card balance

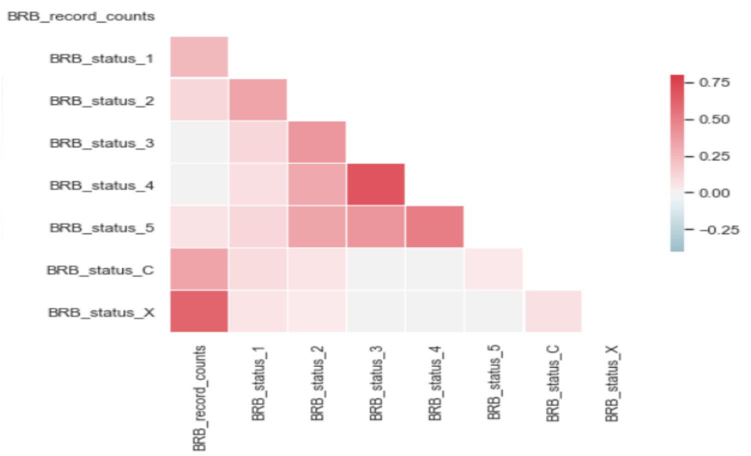


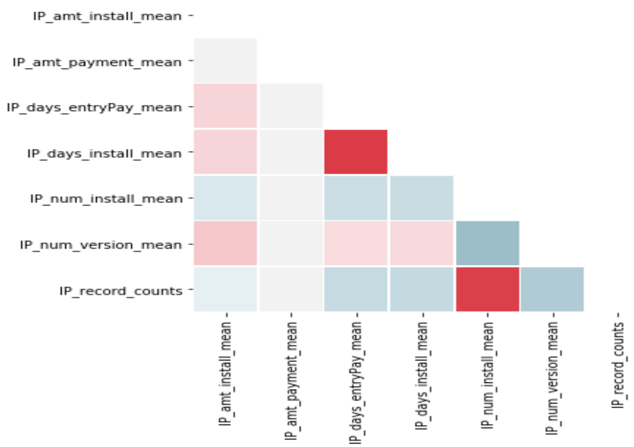
- **Total receivables and mean balance: moderately strong correlation with the coefficient of 0.67**

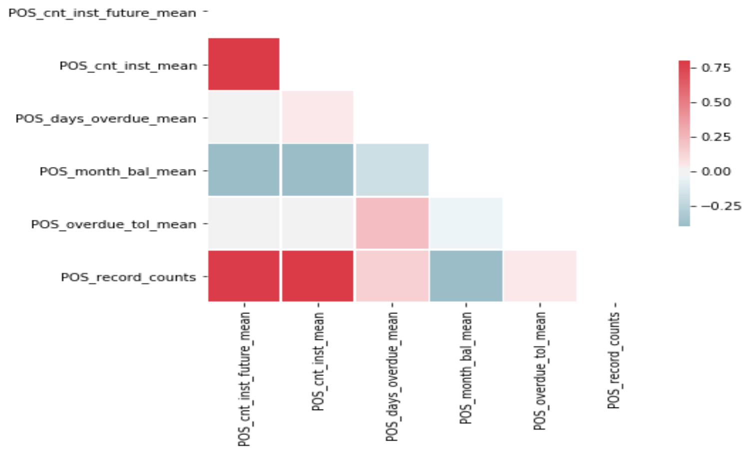
Other historical data

- **(1) Bureau; (2) bureau balance; (3) instalment; (4) POS balance**



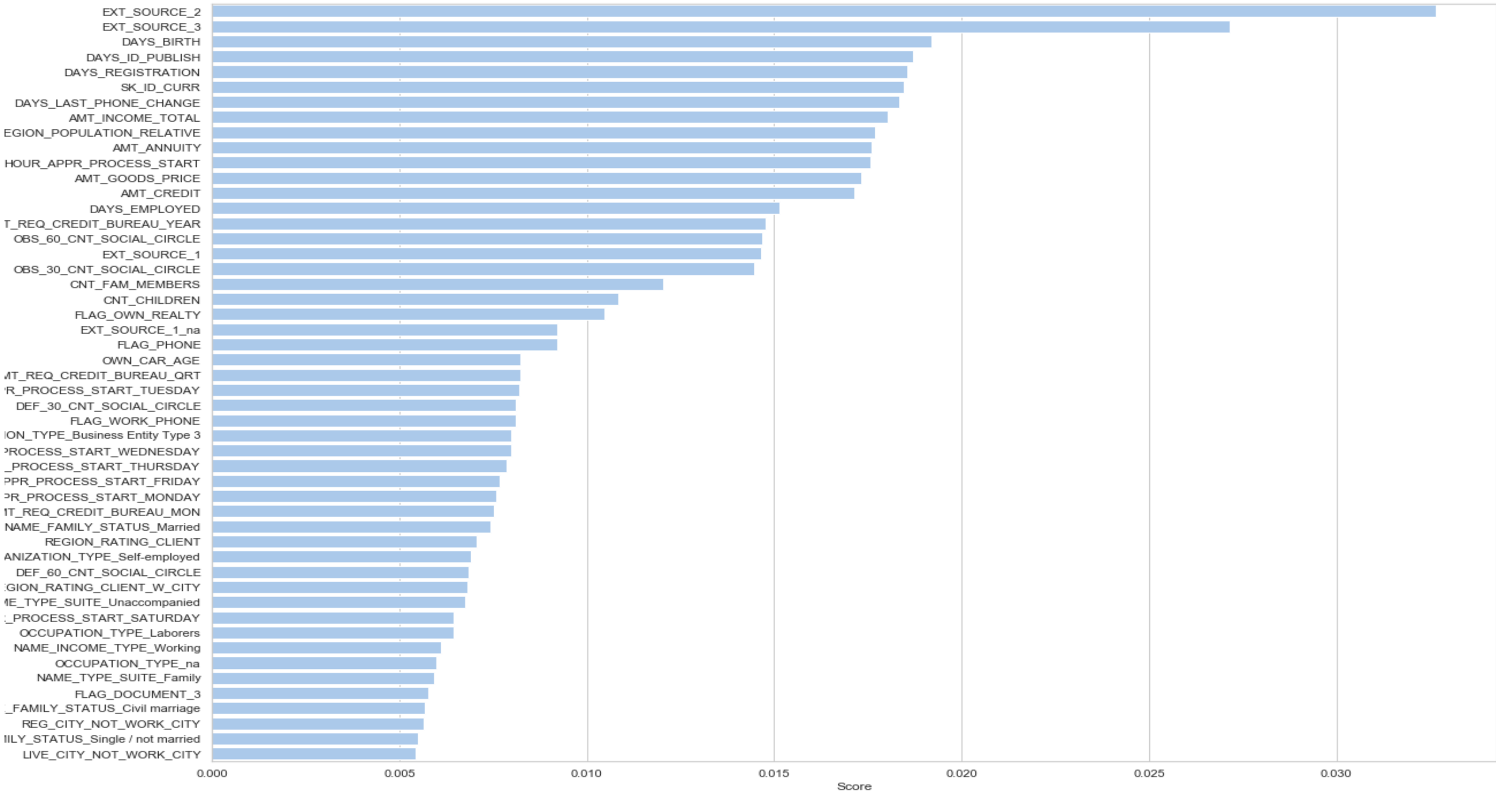






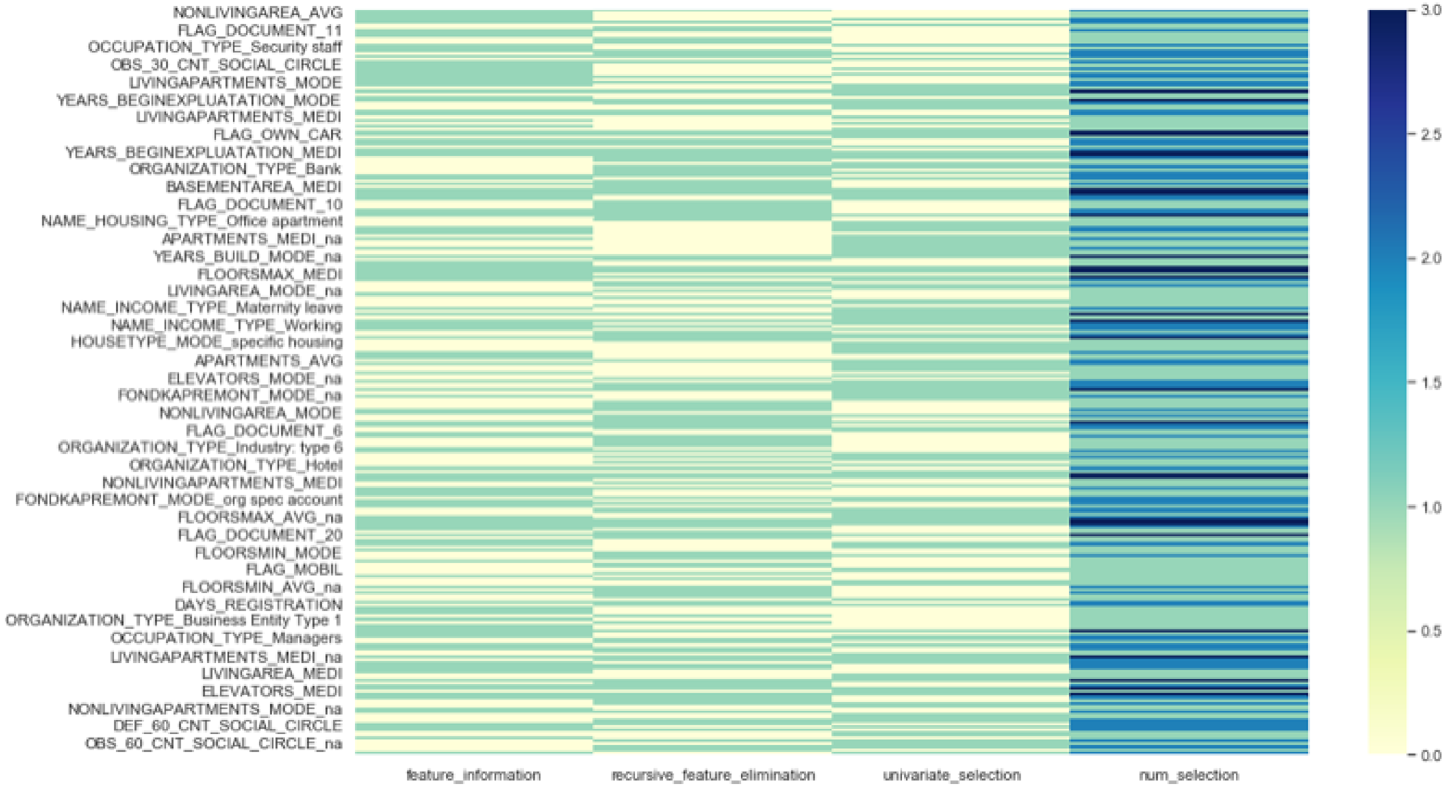
- **Entry pay mean days and Install mean days: r = 0.99**

Feature engineering



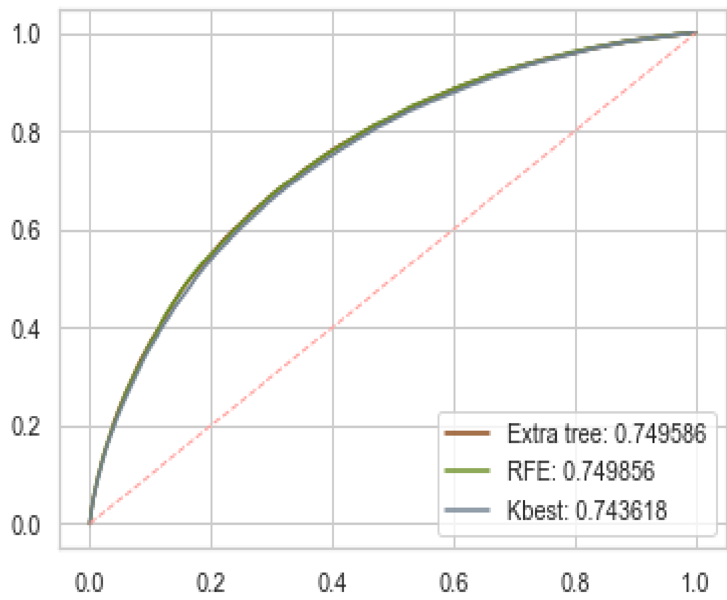
- **To 50 features from extra trees classifier**
- **Highest score: 0.032 (EXT_SOURCE_2)**
- **Second score: 0.028 (EXT_SOURCE_3)**
- **Select 137 features scoring greater than 0.002**

Three algorithms for feature engineering



- **Modelling for application data only**
- **Extra trees classifier, recursive feature elimination and k-best**
- **Total number of 257 features chosen from at least one algorithm**
- **Thirty one features chosen commonly in all three methods**

AUC scores for three algorithms



- **Hold 30% of data for the test and use 70% of data for the test**
- **Logistic regression algorithm**
- **AUC score: 0.74972 for Extra Trees Classifier → Use this**
- **AUC score: 0.74998 for REF**
- **AUC score: 0.74363 for K Best**

Parameter tuning for logistic regression

- **Modelling for all data together**
- **Tuning: penalty, tolerance and regularization parameter C**
- **Penalty range: (l1, l2)**
- **Tolerance range: (1e-3, 1e-4, 1e-5)**
- **C range: (0.01, 1, 100)**
- **Parameter tuning method: Grid search CV 5**
- **Steps: logReg 1 → pram tuning 1 → logReg 2 → feature selection 1 → pram tuning 2 → logReg 3 → feature selection 2 → pram tuning 3 → logReg 4**
- **Four AUC scores from four logistic regression models**

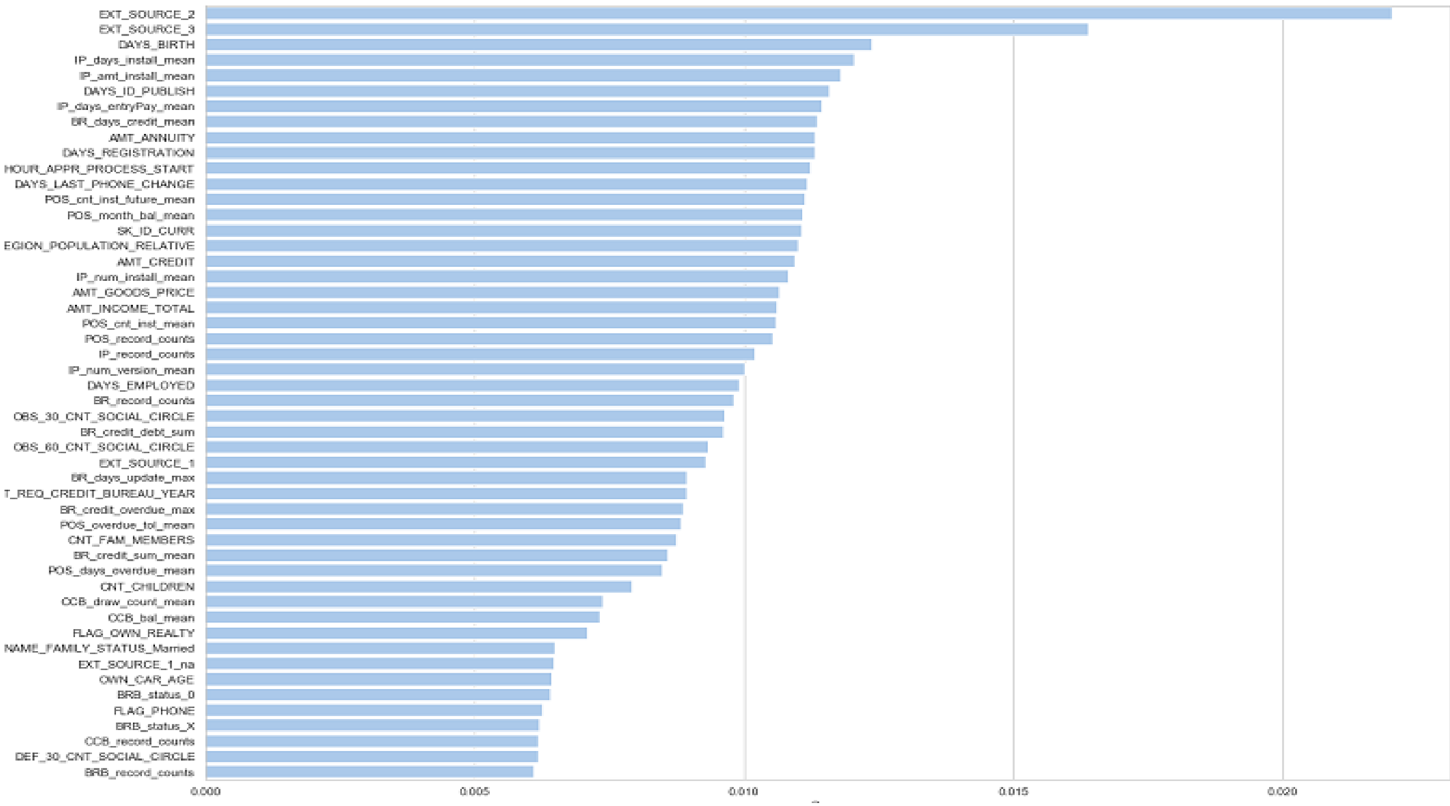
Hyper-parameters

- **Hyper-parameter tuning outcomes**

Tuning	Penalty	Tolerance	C
Tuning 1	L2	1e-5	100
Tuning 2	L2	1e-5	100
Tuning 3	**L1**	1e-5	100

- **Only difference in penalty at the third tuning**

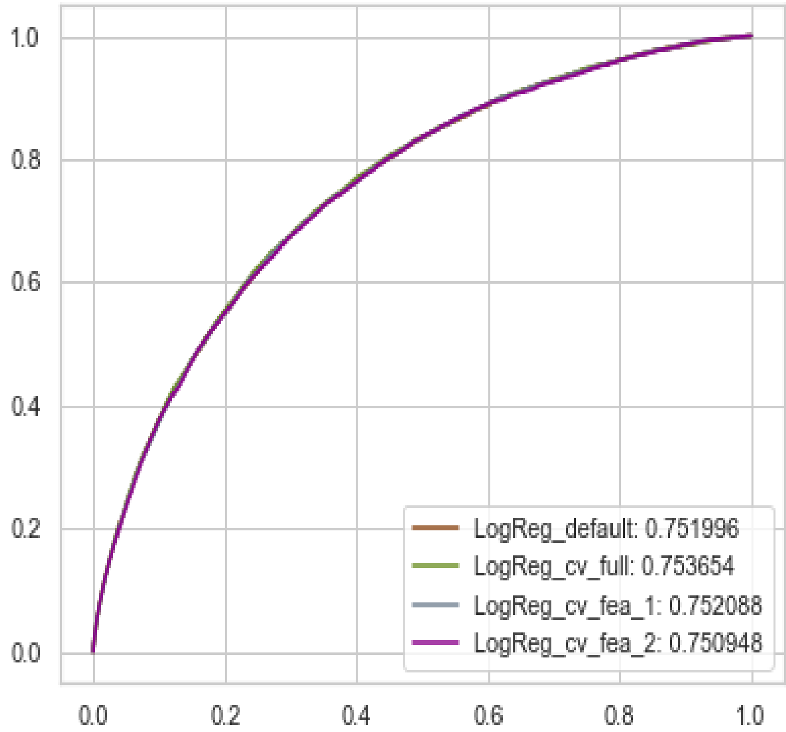
Feature selection



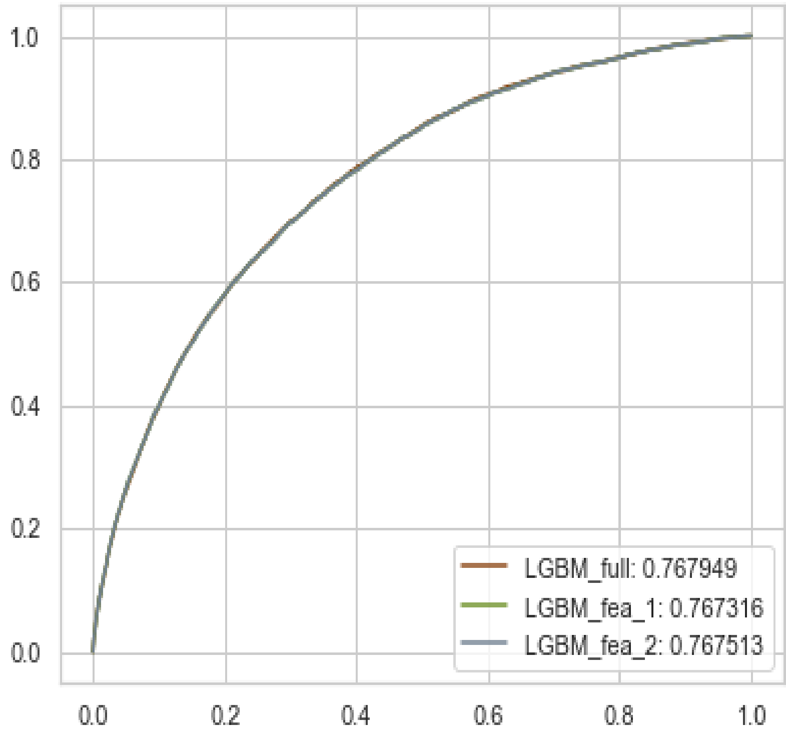
- **Top 50 features from extra trees classifier**
- **Top score: 0.023 for EXT_SOURCE_2**
- **Second score: 0.017 for EXT_SOURCE_3**
- **Same top two as the feature engineering for application data**

ROC curves for models

Logistic regression



LGBM



AUC score table for models

Model	Number of features	Score
Logistic: default	328	0.751996
Logistic: cv5 for all features	328	0.753654
Logistic: cv5 for 206 features	206	0.752088
Logistic: cv5 for 152 features	152	0.750948
LGBM: all features	328	0.767949
LGBM: 206 features	206	0.767316
LFBM: 152 features	152	0.767513

Discussion

- Project of the predictive modelling for the loan application data
- Serious issue in missing values in the application dataset
- Implementation of missingness indicators
- Aggregate historical data and merged with the application
- Three different feature engineerings for logReg on application data
- Extra trees classifier derive the competitive outcome
- Hyper-parameter tuning of logReg for whole data
- LGBM models for different feature selections
- Best AUC score: 0.76795 at the LGBM model with the full features

Limitations

- Lack of expertise on loan business
- Could not apply more sophisticated aggregation methods such as timewise weight or specified missing imputation methods
- Did not apply the hyper-parameter tuning for the LGBM
- Large number of parameters and computationally expensive
- Another project for these in the future

References

[1] https://www.kaggle.com/c/home-credit-default-risk/data

[2] https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html