# Capstone Project Machine Learning Engineer Nanodegree	Andrii Stelmashenko February 13th, 2018

Definition

Project Overview

Loans are very popular in US. People take loans for education, property, car, etc. The definition of loan is:

A loan is money, property or other material goods that is given to another party in exchange for future repayment of the loan value amount along with interest or other finance charges [1].There is a risk that a lender may not receive given credit back, it's called 'Credit Risk' [2]. Lenders, of course, want to minimize credit risks, for that they calculate 'Credit Score' [3]. A credit score is a number representing the creditworthiness of a person, the likelihood that person will pay his or her debts. Credit score is composed from several components related to a person, one of major components is 'Credit History'. A credit history is a record of a borrower's responsible repayment of debts.

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. It's possible to use alternative data, including telco and transactional information, to predict their clients' repayment abilities, which can help people to get loans from trustworthy lenders.

Problem Statement

There is a company Home Credit which works with the category of unbanked population. They have a set of data: client's previous credits provided by other financial institutions that were reported to Credit Bureau; Monthly balances of previous credits in Credit Bureau; monthly snapshots of credit card balances, cash loans applicant has/had with Home Credit; previous applications for Home Credit. Also they have a historical target score for existing customers, they want to predict for new applicants. The goal of this project is to predict probability that a person will pay it's debts based on related to the person financial information, some kind of alternative to 'Credit Score'.

Dataset is provided by competition organizer - HomeCredit company, and it is hosted on kaggle platform. There are seven different sources of data:

- application_train/test: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
- bureau: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
- bureau_balance: monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
- previous_application: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV.
- POS_CASH_BALANCE: monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
- credit_card_balance: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
- installments_payment: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

HomeCredit_columns_description.csv file contains descriptions for the columns in the various data files. There are 221 row in the file with human-readable description for each column.

Metrics

Submissions are evaluated on area under the ROC curve [4] between the predicted probability and the observed target. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The evaluation logic is the next: submitted results are actually two probability distributions of two classes - 0: the loan was repaid or 1: the loan was not repaid. ROC curve will show how we did that split, it visualizes all possible classification thresholds. AUC represents the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation.

The result will help HomeCredit business to decide if they want e.g. to minimize False Positive Rate or maximize True Positive Rate.

Analysis

Data Exploration

Dataset is provided by competition organizer.

HomeCredit_columns_description.csv file contains descriptions for the columns in the various data files. There are 221 row in the file with human-readable description for each column. Relations between files are provided as a diagram:

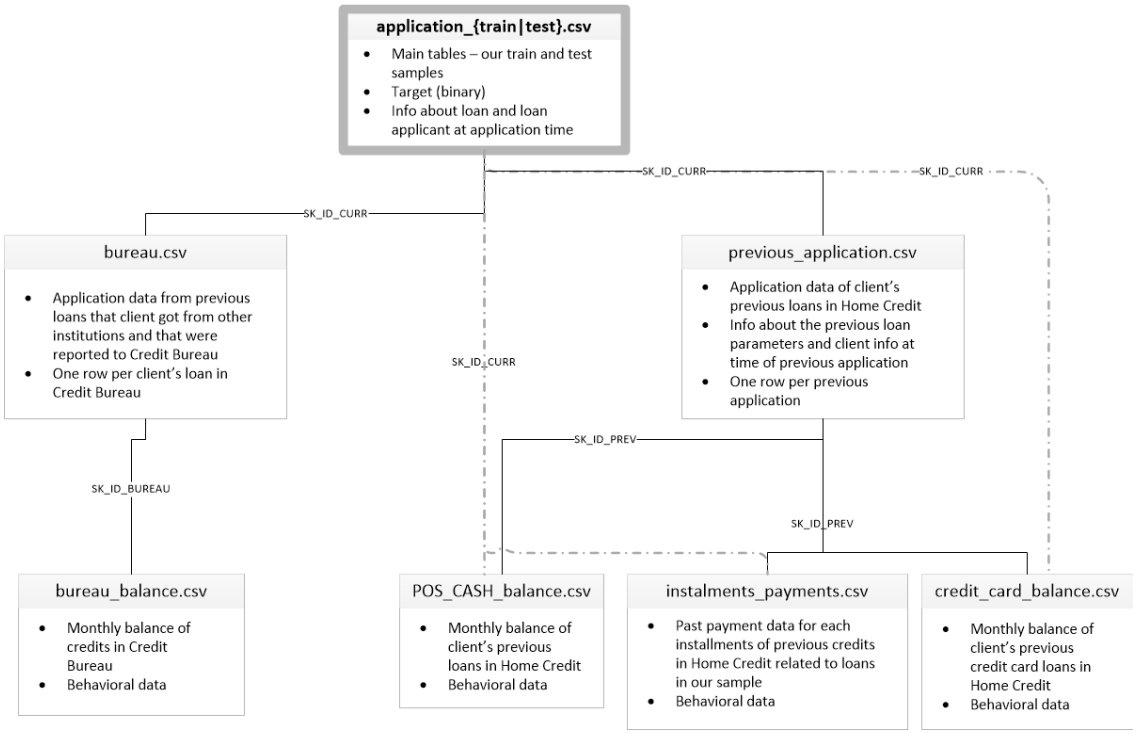

Figure 1 – Input data relations

Data files are split, later all these files should be concatenated using techniques like SQL Left Join[5] into single csv file with 221 column. Later in text analysis is provided only of some of columns which are discovered to be interesting, general analysis will be done only on one column as an example, other similar columns will be skipped to reduce document size.

All columns are of one of three data types: int, float, object. These types correspond to logical data type of each column: continuous, numerical discrete and categorical. Categorical columns are represented by object and int types. Let's count unique values:

```
app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)
```

Partial output:

```
NAME_CONTRACT_TYPE              2
NAME_EDUCATION_TYPE             5
WEEKDAY_APPR_PROCESS_START      7
ORGANIZATION_TYPE              58
```

Some columns have only 2 unique values and some have more - 58. Most algorithms work only with numerical values, so it is necessary to encode string values to numerical. Two most used techniques for that are Label Encoding and One-Hot-Encoding.

Columns of continuous type should be analyzed on anomalies. anomaly is the deviation in a quantity from its expected value[6]. Since it's often unclear what is 'expected value' there is number of statistical methods how to detect anomaly or outlier, Tukey's mdethod[7] is one of them. It is based on interquartile range (IQR), the range is the next:

```
[Q1 - k * (Q3 - Q1), Q3 + k * (Q3 - Q1)]
```

Where `k=1.5` indicates an outlier, `k=3` means value is "far out".

Let's take AMT_INCOME_TOTAL column and apply IQR, the column has 3014 values which are far out from the rest. Executing outliers search for all continuous columns gives 3381 outlier when `k=3` and 25350 when `k=1.5`. AMT_INCOME_TOTAL contains most outliers. The column holds income of a client. These data points are very important because if client's income is big then most likely it will take big loan, if we look at such clients (see Data_Exploration.ipynb):

```
app_train.loc[income_total_outliers]['TARGET'].astype(int).value_counts()
0    2852
1     162
```

There is portion of rich clients which do not return loans back.

It depends on algorithm chosen if removing outliers is necessary, some algorithms are sensitive to outliers and some are not. In case of this projects outliers may be very important and it is better to chose algorithms which are not sensitive to outliers.

Missing values is another problem in real world projects. There are a lot of columns where missing values percentage goes up to 60:

Column Name	NaN %	NaN Count
COMMONAREA_AVG	69.872297	214865
NONLIVINGAPARTMENTS_MODE	69.432963	213514
NONLIVINGAPARTMENTS_MEDI	69.432963	213514
...		

Table 1 – Missing values

Having missing values in a dataset can cause errors with some machine learning algorithms. There are number of options what to do with missing values: drop rows or event whole columns where there are to many missing values, impute missing values, e.g. use mean or zeros.

Research on numerical columns using Five Number summary [9] showed that DAYS_EMPLOYED column has group of abnormal value 365243 which is how many days before the application the person started current employment. This values means a person worked 1000 year before application, it seems like this number is sort of flag information is unknown or is not filled by some reason.

Exploratory Visualization

Target column is binary, it takes either 0 or 1 values. Let's look it histogram plot.

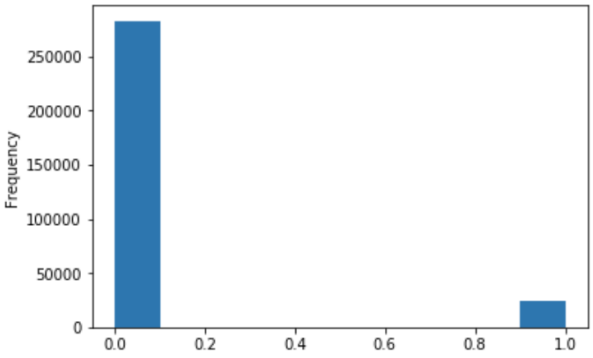

Figure 2 – Target distribution

From the fig. 2 it is seen that there is imbalanced class problem[8].

Missing values, was mentioned already above, can be visualized using data-dense matrix:

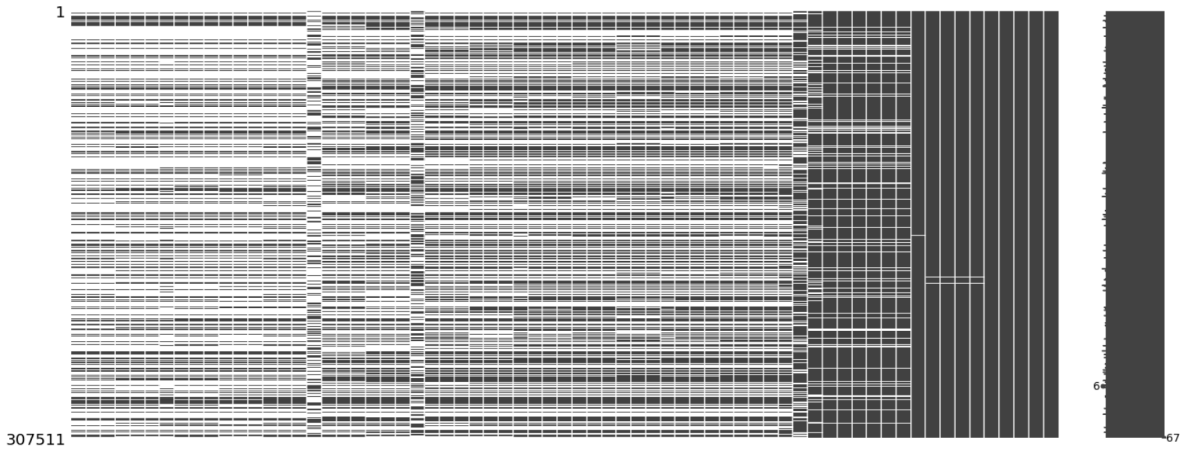

Figure 3 – Missing values matrix

It shows that most rows have missing values in a lot of columns at the same time (see horizontal blank lines). Columns are sorted from the biggest portion of missing values to the smallest:

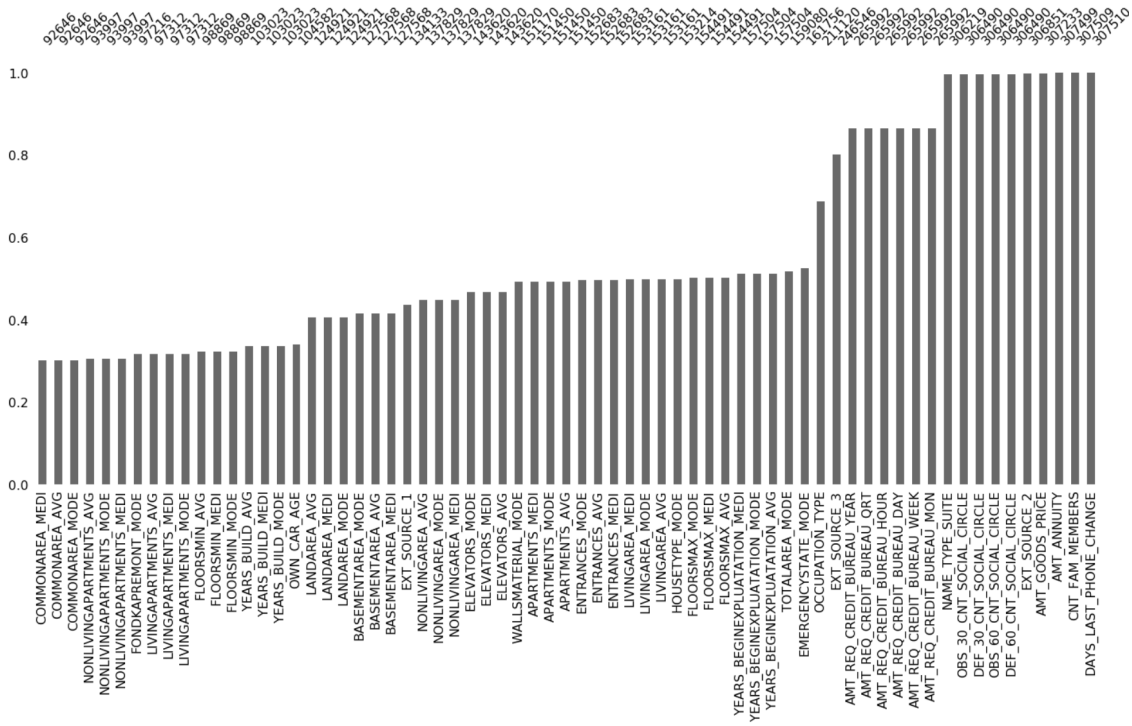

Figure 4 – Missing values barplot

The bar plot shows which columns have the biggest portion of missing values, which are mostly related to apartment/house. Certain algorithms require missing values to be filled in and some others can work with missing values. The decision what to do with missing values will be made after algorithm has been chosen.

There is certain amount of correlations exist in the data set (see Figure 5 below). On the below heatmap are shown only columns with correlations. There are columns describing apartment/house aspects and they are represented as _AVG, _MODE and _MEDI, which are average, modus and median. Correlation value is close or equal to 1 which is very strong correlation type and may considered as duplicates. These columns are candidate for removal e.g. using principal components analysis technique or even, just dropping the columns _MEDI and _MODE leaving only _AVG ones.

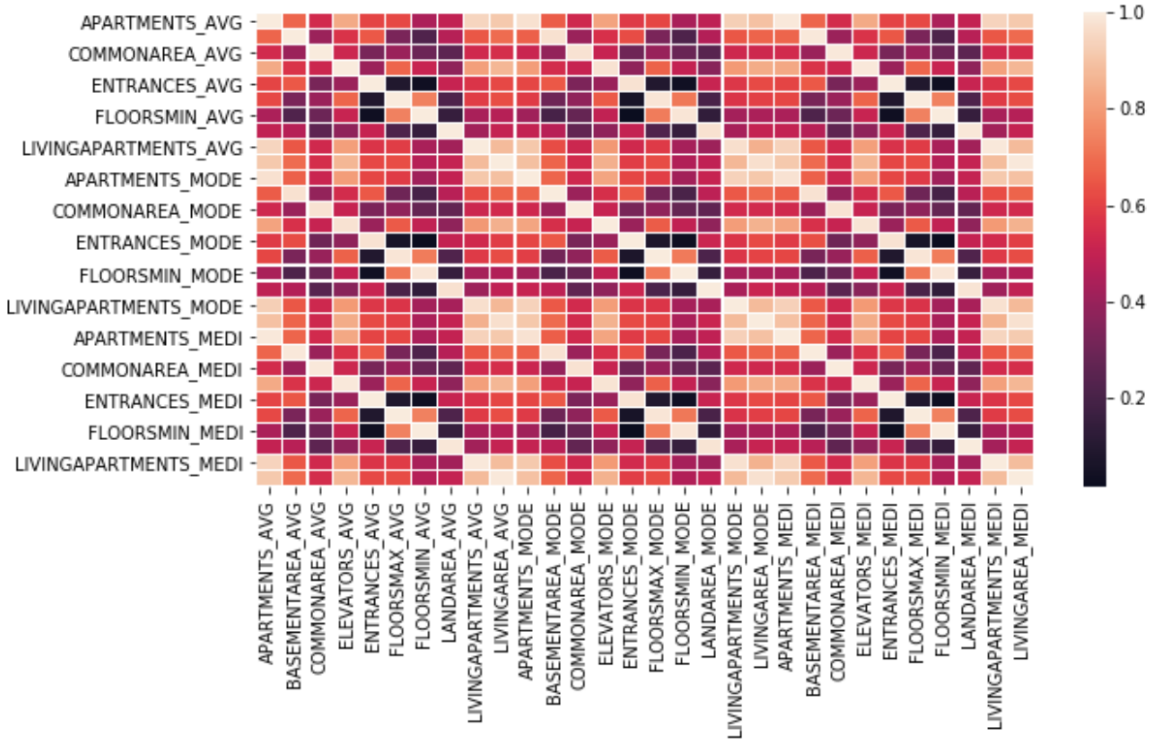

Figure 5 – Correlations

One more candidate for visualization is DAYS_EMPLOYED column, it's worth to look at hist plot with and without outliers:

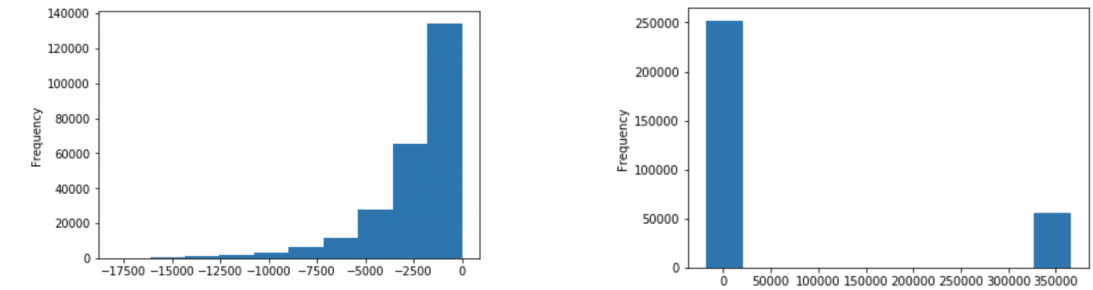

Figure 6 – Days employed column with and with out abnormalies

With out abnormal points distribution is more "gaussian", many algorithms make assumption that data distribution of numerical columns is "gaussian", which may affect model performance.

It's also useful to visualize outliers to see how far from other values they are, below is boxplot of OWN_CAR_AGE column:

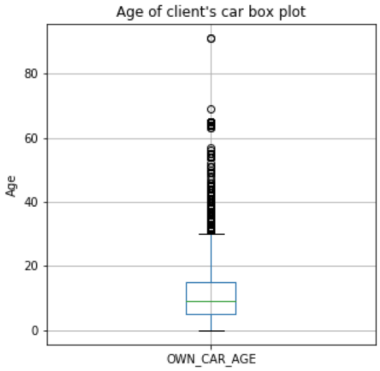

Figure 7 – Outliers visualization

From the plot it is possible to evaluate how many outliers are there outside of IQR.

Algorithms and Techniques

Gradient Boosting Machine (LightGBM [10]) is a good choice for this type of data. LightGBM: gradient boosting machines proved to perform good on high dimensional datasets with mixed categorical and numerical features, it can natively handle categorical features and sparse datasets (see [10] section 4). It is very fast, it uses techniques Gradient-based One-Side Sampling and Exclusive Feature Bundling to deal with large number of data instances and large number of features respectively. It supports parallel execution out of the box [11]. Tree based models are interpretable [12], [13] which is strong side.

Continuous values handling is one of the weaknesses of tree based models, LGBM uses discretization technique which leads to information loss, it is not critical in most cases.

Benchmark

A naive base line would be a random guess, taking into account we use ROC as a metric, random guess will give us a straight line because true positive rate and false positive rate will be the same. In addition to the straight line it'd be nice to add a simple model result to compare to, e.g. Logistic Regression [17].

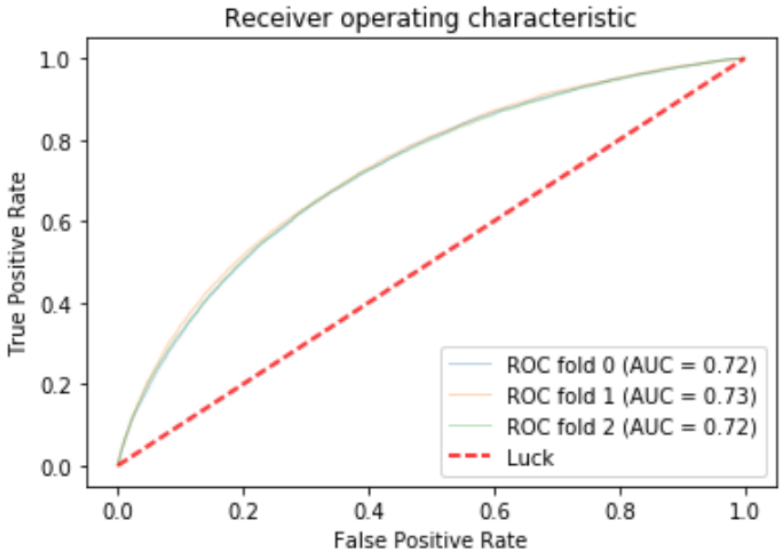

Figure 8 – Base Line

Logistic regression gives 0.72 AUC value, it is obtained based on the data, missing values are fulfilled with median imputation strategy. Red dash line corresponds to random guess result.

Methodology

Data Preprocessing

Another important thing is to use all provided data. Additional files provided represent applicants additional data, and relates as one-to-many to application train/test files rows. A simple strategy of merging these additional data would be to do a 'left join' [5], however relation is one-to-many does not allow us to do that without duplicating rows. A solution to that can be averaging values grouped per applicant for numerical values and taking most frequent values for categorical features, this is what is done with provided files. This can be considered as Feature Engineering and it's hard to predict how it will affect algorithm performance. For below code examples see Data_Merging.ipynb notebook.

Numerical feature aggregation example, load bureau data

```
bureau = pd.read_csv('input/bureau.csv.zip')
bureau_by_skid = bureau.groupby('SK_ID_CURR')
avg_bureau = bureau_by_skid.mean()
# merge bureau data and application data
app_train = app_train.merge(avg_bureau, how='left', on='SK_ID_CURR')
```

For categorical values it's possible to find most frequent value:

```
credit_card_balance = pd.read_csv('input/credit_card_balance.csv.zip')
max_status = credit_card_balance.groupby('SK_ID_CURR').agg(most_frequent)
credit_card_balance['MAX_STATUS'] = max_status['NAME_CONTRACT_STATUS']
app_train = app_train.merge(credit_card_balance, how='left', on='SK_ID_CURR')
```

This way all numerical and categorical features of additionally joined tables were processed. Additionally conflicting column name were renamed:

```
cols_rename = {'MONTHS_BALANCE': 'CC_MONTHS_BALANCE',
               'NUNIQUE_STATUS': 'CC_NUNIQUE_STATUS',
               'SK_DPD': 'CC_SK_DPD',
               'SK_DPD_DEF': 'CC_SK_DPD_DEF'}
credit_card_balance = credit_card_balance.rename(index=str, columns=cols_rename)
```

Few more new features added are counts, e.g. count of previous applications for each applicant, this gives additional information which otherwise would be lost during averaging and left join operations:

```
cnt_previous_app = previous_application.groupby('SK_ID_CURR').count()
avg_previous_app['NUM_APPS'] = cnt_previous_app['SK_ID_PREV']
```

LightGBM can natively process missing values [11] the same way as it handles sparse matrices [18], so there is no need to impute missing values and even do e.g. One-Hot-Encoding just specify categorical features column names. The thing was done translating `object` column types to `categorical`

```
cat_features = train.select_dtypes('object').columns.tolist()
for col in cat_features:
    train[col] = train[col].astype('category')
    test[col] = test[col].astype('category')
```

Redundant columns found on Data Exploration stage are removed – columns with _MEDI and _MODE suffixes, it reduces data dimensionality from 197 to 177 columns.

LightGBM is not sensitive to outliers because it uses histogram-based algorithms, which buckets continuous feature values into discrete bins.

Implementation

LightGBM implementation has many parameters [19]. There most important are:

- num_leaves – main parameter to control complexity. Recommended to to less then 2^(max_depth) to prevent overfitting
- min_data_in_leaf – parameter to prevent overfitting, it's value depends on train samples and num_leaves. Setting it larger prevents trees to grow too deep which may cau under-fitting
- max_depth – limiting tree depth, if there is not limit (-1) it may cause overfitting
- max_bin – max number of bins that feature values will be bucketed in
- l1 and l2 regularization

This is a GBDT algorithm and it tends to overfitting if it is not restricted on trees growth. There is a set of recommendation on the official documentation how to deal with overfitting through parameters few of them were applied:

- l1 and l2 regularization parameters
- max_depth restricted
- num_leaves decreased
- data set is big enough, it allows to train on small learning rate

Refinement

From data exploration there is a fact that this is imbalanced data set, so it is important to use stratified splits into training and validation folds. ‘Stratified’ here means that while splitting data set proportion on 0 and 1 will be preserved in training and validation sets. Below code example are from Lightgbm.ipynb notebook.

First result on application_train/test data sets was auc=0.746, parameters:

```
n_estimators=10000,
objective='binary',
class_weight='balanced',
learning_rate=0.0003,
num_leaves=31,
max_depth=-1,
reg_alpha=3,
reg_lambda=5
```

After merging other files to the application_train/test and on the same hyper parameters result improved to auc=0.767.

Parameters tuning is very important stage to find proper ones for the data set. Random search for that is proper tool in sklearn library. It is proved that random search may give better parameters then grid search [20]. Out of the box implementation did not work (by some reason kernel died), so custom implementation was done based on sklearn ParameterSampler and Kfold to split train and validation sets.

Parameters to try look like this:

```
params = {
    'learning_rate': [0.01, 0.1],
    'n_estimators': [8000, 16000],
    'num_leaves': [16, 32],
    'max_depth': [-1, 4, 7],
    'min_child_samples': [10, 20, 40],
    'reg_alpha': [0.1, 0.5, 1.0],
    'reg_lambda': [0.1, 0.5, 1.0],
    'min_data_in_leaf': [8, 16, 32],
    'max_bin': [128, 256]
}
```

This process of searching parameters has been run iteratively, at start big learning rate was chosen to understand which hyper parameters perform good and then learning rate was decreased to fine tune parameters. It ended up with the next parameters:

```
n_estimators=10000,
objective='binary',
class_weight='balanced',
learning_rate=0.0003,
min_child_samples=160
num_leaves=31,
max_depth=7,
reg_alpha=2.0,
reg_lambda=2.0
```

This improved results from auc=0.767 to acu=0.773 on the test set.

Results

Model Evaluation and Validation

Final parameters are outlined in the Refinement section. To make sure algorithm is robust train and validation sets were used. Train data sets was split into Train and Validation using K-Fold technique. Train data set is to train algorithm parameters, validation data set is to find hyper-parameters. And finally test unseen data was used to evaluate algorithm performance. This approach makes sure algorithm is robust and has good generalization.

To check if model generalizes well enough plot is build iterations on X axis and AUC on Y axis for validation and train splits.

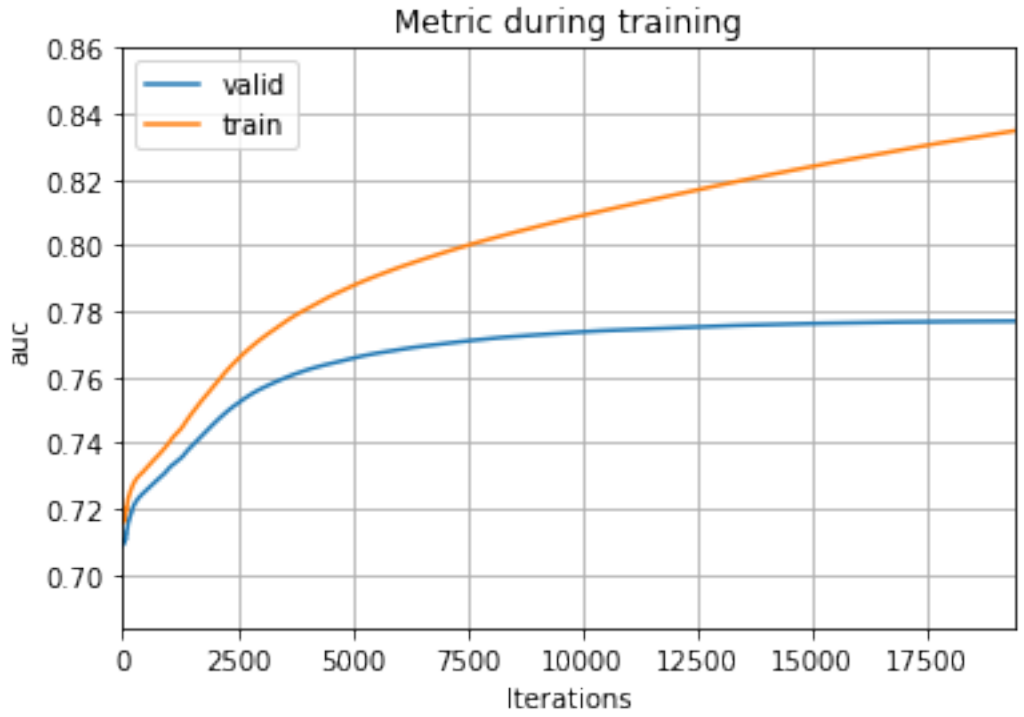

Figure 9 – Validation vs Train AUC metric

Justification

Base line solution result was auc=0.72. Final algorithm result is auc=0.773 which is not that big but still good enough improvement, taking into account that each additional tenth of improvement took more and more time finding more appropriate parameters and training time using small learning rate. And it is much better then random guess result auc=0.5.

Conclusion

Free-From Visualization

Lightgbm algorithm gives one more useful thing feature importance. After algorithm is run trained it’s possible to plot it:

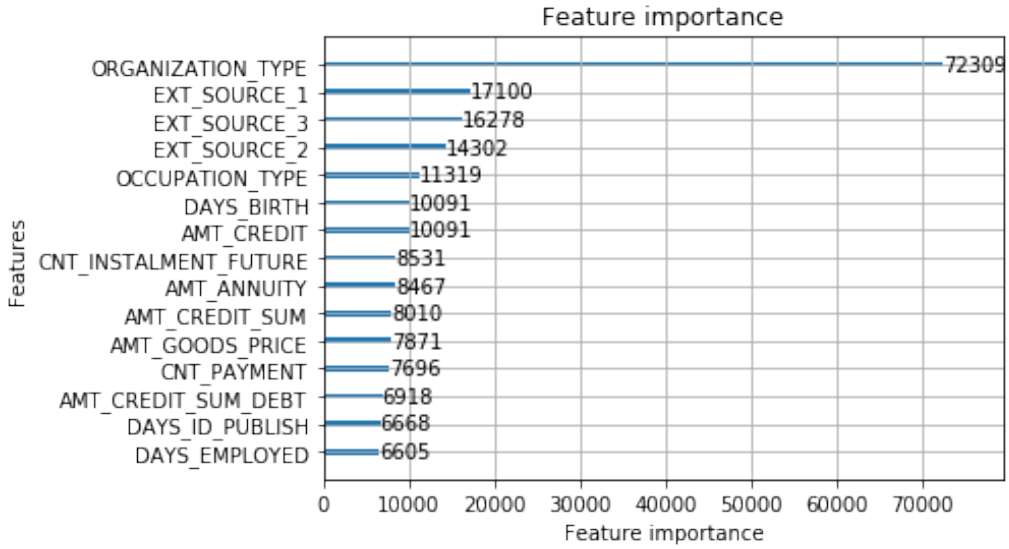

Figure 10 – Feature Importance

From the plot it is seen that organization type is the most important feature with much higher importance value then other features.

Reflection

The result of the project is important to the HomeCredit business. The final model will produce a probability that a new or returned applicant will return credit back, based on just knowledge about the applicant and previous experience. This can be used to improve the process of giving loans with lower risk, which in long term will increase company benefit and the same time additional knowledge may be used to improve process of providing loans for people with low probability of returning credit in some way more people who needs loan will receive it which may make somebody's life better. And this is complex part of this project because it touches ethics of what HomeCredit company does, it may affect certain people life.

From technical perspective the complexity of this project in provided data, it has a lot of issues to be addressed: missing values, a lot of additional data which do not relate as one-to-one to a row and it is not clear if additional data help or make results worse.

Improvement

There is a room to improve the project. Fine tuning parameters through random search is limited only by time, this means it's very hard to find extremely optimal parameters, it's highly probably that final parameters will be some local optimal.

Another way to improve performance is feature engineering. Using provided additional files with data it's possible to create new features which do not correlate with existing features, these new features is additional information which may improve performance.

Ensemble of models will most likely improve results as well. It's possible to combine results of heterogeneous models e.g. using Stacking [16]. This very time and resource consuming because requires training another algorithm and hyper-parameters tuning.

Resources

1. Investopedia – Loan (https://www.investopedia.com/terms/l/loan.asp)
2. Wikipedia - Credit Risk (https://en.wikipedia.org/wiki/Credit_risk)
3. Wikipedia - Credit Score (https://en.wikipedia.org/wiki/Credit_score_in_the_United_States)
4. Wikipedia - ROC curve (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)
5. Wikipedia - LeftJoin (https://en.wikipedia.org/wiki/Join_(SQL))
6. Wikipedia – Anomaly (https://en.wikipedia.org/wiki/Anomaly_(natural_sciences))
7. Wikipedia – Tukey's fences (https://en.wikipedia.org/wiki/Outlier#Tukey's_fences)
8. Archiv.org – Imbalanced Class Problem (https://arxiv.org/pdf/1305.1707.pdf)
9. Wikipedia – Five Number summary (https://en.wikipedia.org/wiki/Five-number_summary)
10. LightGBM (https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf)
11. Lightgbm – Features (http://lightgbm.readthedocs.io/en/latest/Features.html)
12. GBDT interpretability (https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27)
13. LightGBM interpreter tool (https://github.com/slundberg/shap)
14. Embeddings (https://developers.google.com/machine-learning/crash-course/embeddings/categorical-input-data)
15. Hashing (https://alex.smola.org/papers/2009/Weinbergeretal09.pdf)
16. Wikipedia – Stacking Ensembel (https://en.wikipedia.org/wiki/Ensemble_learning#Stacking)
17. Wikipedia – Logistic Regression (https://en.wikipedia.org/wiki/Logistic_regression)
18. Lightgbm – Categorical Features Support (https://www.researchgate.net/publication/242580910_On_Grouping_for_Maximum_Homogeneity)
19. Lightgbm Parameters (http://lightgbm.readthedocs.io/en/latest/Parameters.html)
20. Random Search (http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf)