Credit Score Prediction

Presented by
Artem Ramus

Introduction and Background

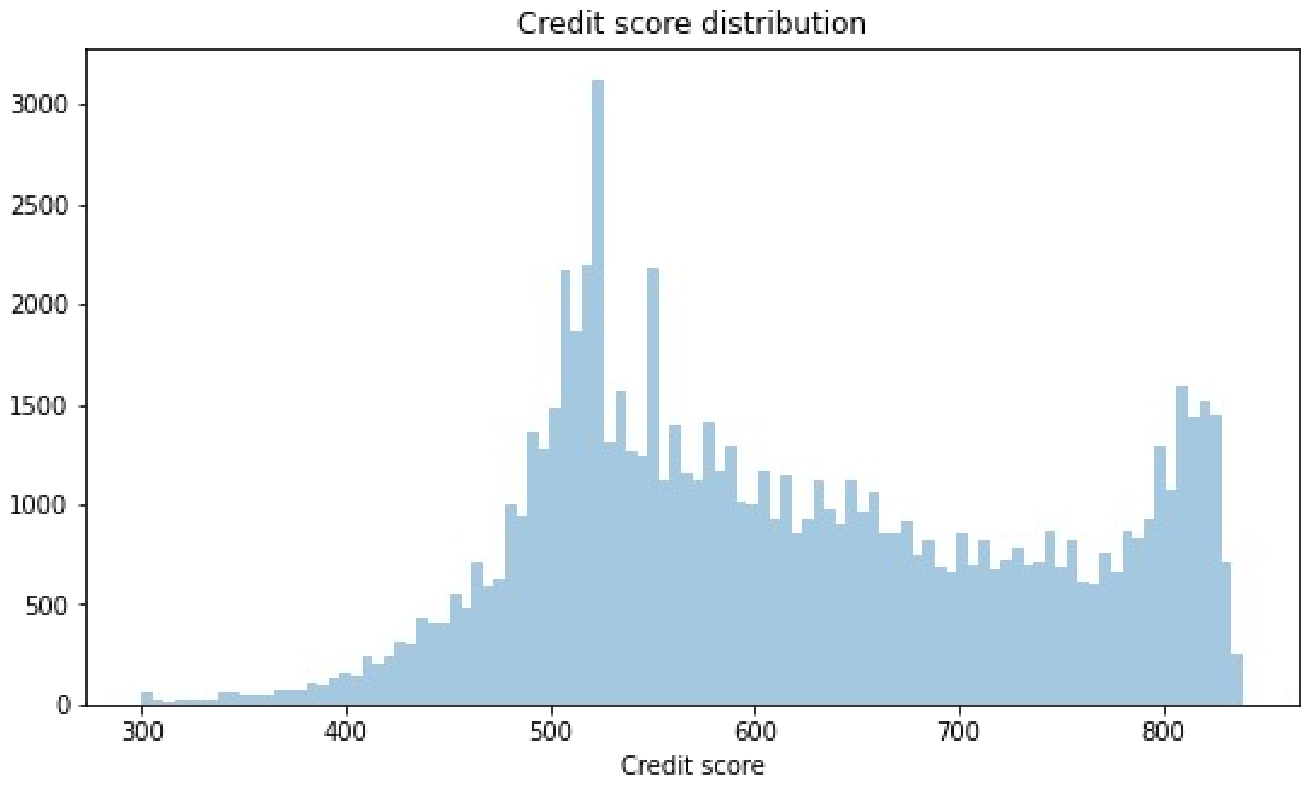
The goal of this project is to predict a credit score from 304 attributes.

About the data from the publisher
We provide you with a data set in CSV format.
The data set contains 8,000 train instances and 2000 test instance
There are 304 input features, labeled x001 to x304.
The target variable is labeled y.

Task: Create a model to predict the target variable y.
A report - A Power point presentation
Any custom code you used
Instructions for me to run your model on a separate data set

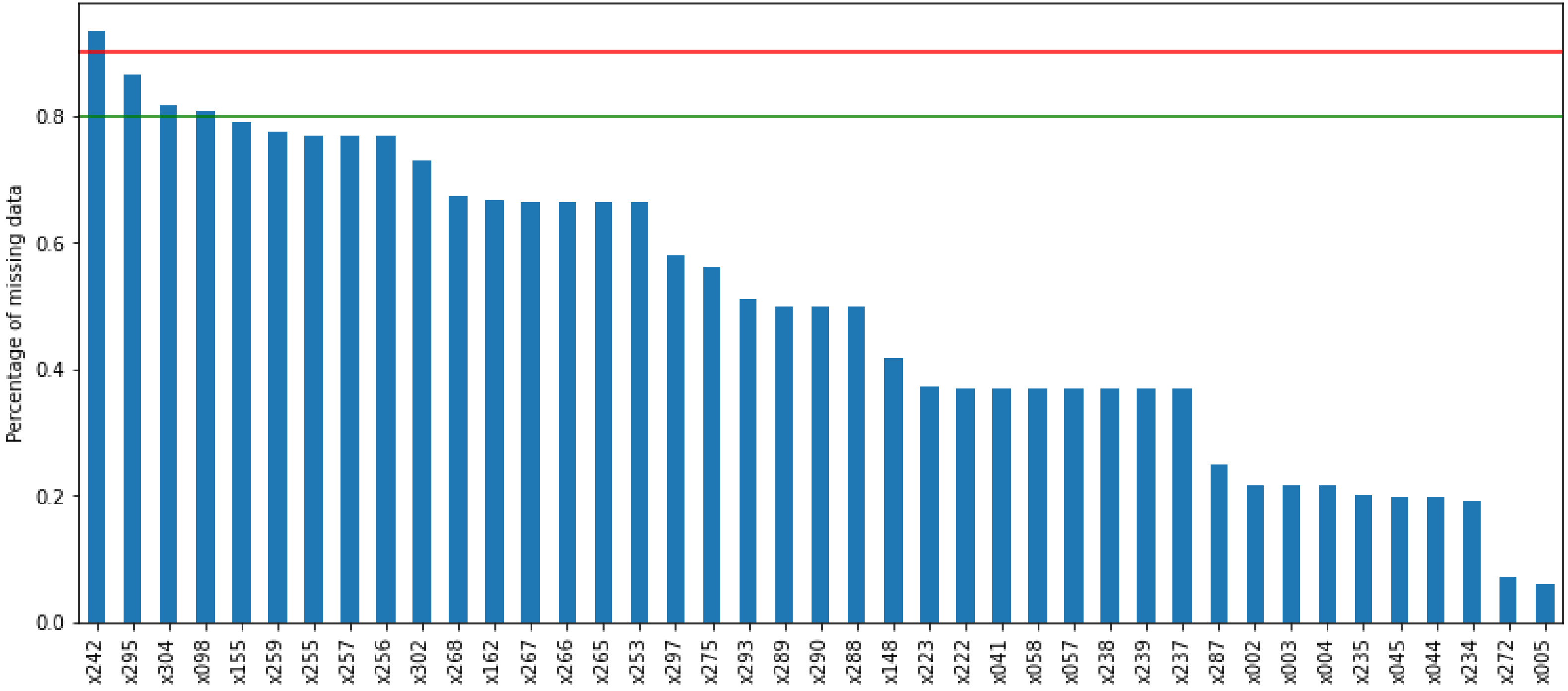
Link to the data set at Kaggle:
https://www.kaggle.com/prasy46/credit-score-prediction

Exploratory Data Analysis – target



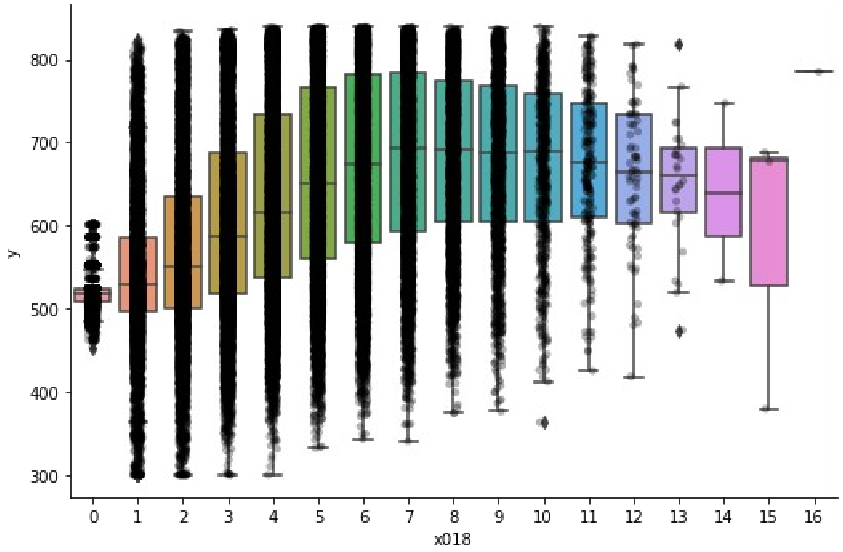
The target distribution has two peaks, at 520 and at 830.

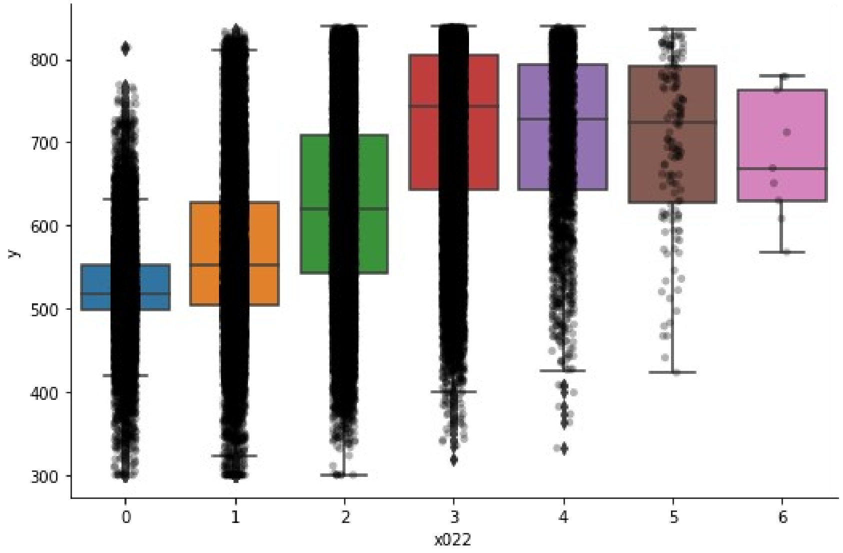
Exploratory Data Analysis – missing values



There are 41 variables with missing values. 4 of them has more than 80% missing values.

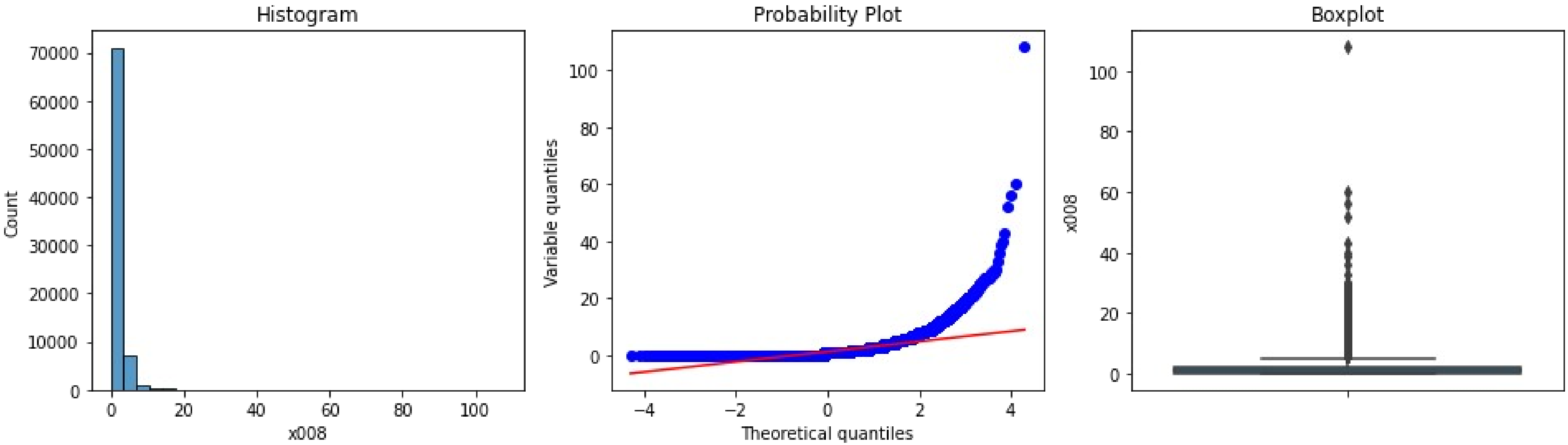
Exploratory Data Analysis – discrete variables





There are 142 variables . 114 of them vary with the target and may be used as predictors.

Exploratory Data Analysis – continuous variables



There are 162 variables . Most variables are skewed or/and have high kurtosis. Some of them have outliers on one or both sides. 5 are highly skewed. Many continuous variables seem slightly to not correlated with the target.

Feature Engineering

- 28 features with indication of missing values are created. This is based on EDA, variables are missing values have influence on the target.
- In discrete variables missing values are replaced with mode.
- In continuous variables missing values are replaced with mean.
- There are 333 engineered features.

Feature Selection

- Based on the EDA, 114 discrete features and all the continuous features that might influence the target are taken as a starting point.

- 9 quasi-constant features with more than 99% of unique values are deleted.

- 144 correlated more than 90% features are deleted

- There are 202 engineered features.

Performance

Model
XGBoost regressor

Cross validation
RMSE: 92.45 %, Standard Deviation: 0.20 %

Evaluation
RMSE train: 1037.0
RMSE test : 1065.4
r2 train : 0.9260
r2 test : 0.9246

Summary

- 202 of 304 features are selected for modeling.
- With the selected features credit scores were predicted with RMSE_train of 1037.0 and RMSE_test of 1065.4. R2_train 0.9260 and R2_test 0.9246.

The end

Thank you for your attention!