Home Loan Default Risk Prediction

(Team - Caesar)

Team Member 1 - Dhruv Awasthi (Point of Contact)
Team Member 2 - Ashok Senapati

Key Features

- Command line interface
- Built as a pipeline
- Modular approach
- Abstract components
- Easily scalable (current size ~24.4kB)
- Log everything
- Exception handling
- Save model for later use
- Single file configuration
- Pickle dump important objects
- Version control system
- No hard coding, can be used for n-number of features and any dataset

Preprocessing Pipeline

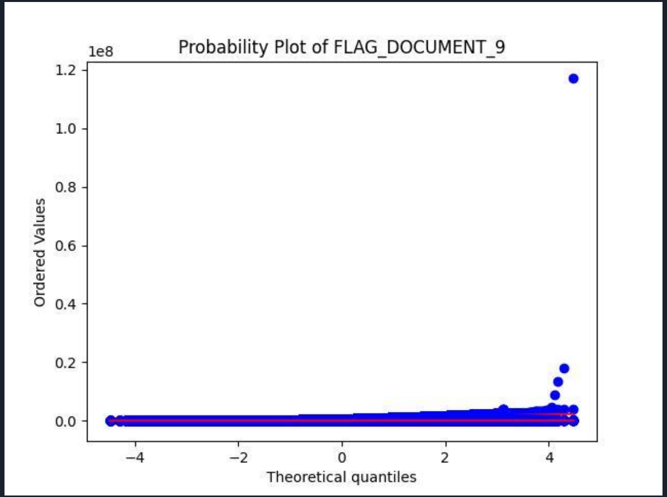
- Removing duplicate rows for training
- Dropping columns with low standard deviation
- Dropping features that do not contribute to the learning
  - Found using correlation matrix
  - Brainstorming
  - Online research
  - Example, `LIVE_CITY_NOT_WORK_CITY`, `REG_CITY_NOT_WORK_CITY`

Preprocessing Pipeline

- Normally Distributed?
  - P-P Plots
  - KS Test
  - SW Test
- The P-P Plots examine the actual cumulative probabilities of your date from that expected from a theoretical normal distribution.
- The KS Test and SW Test examines whether a variable conforms to a predetermined type of distribution (example normal distribution) or whether it differs significantly.
- KS Test is insensitive to minor deviations.

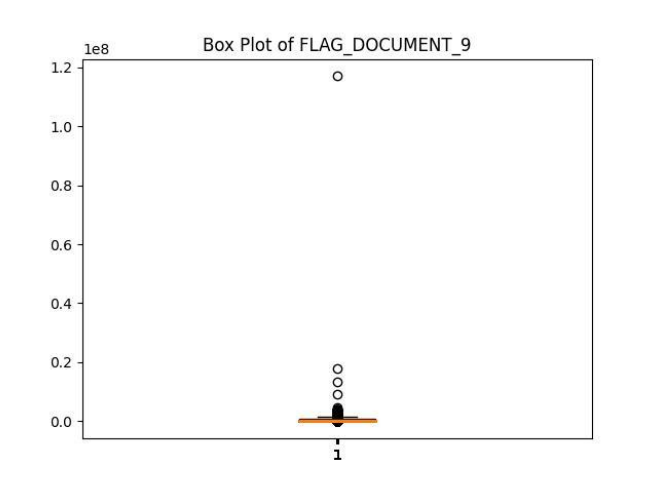
Preprocessing Pipeline

- Outliers
  Causes:
  - Data errors
  - Intentional or motivated misreporting
  - Sampling error or bias
  - Standardisation failures
  - Distributional assumptions



Preprocessing Pipeline

- Identify Outliers
  - Univariate Outliers
    - IQR Percentile
    - Box Plots
    - Trimmed Mean
    - Windsorized Mean
  - Multivariate Outliers
    - Cook's Distance
    - Mahalanobis Distance



Preprocessing Pipeline

- Deal with missing values
- We first need to identify what is the reason for the outliers and the missing values?
    - Missing Completely At Random (MCAR)
    - Missing At Random (MAR)
    - Missing Not At Random (MNAR)
- Random missingness can be problematic from a power perspective but it would not bias the results. However, data missing not at random could potentially be a strong biasing influence.

Preprocessing Pipeline

- Effects of Deletion
  - By deleting samples with missing data, a researcher could be misestimating the population parameters, making replication less likely.
  - If each variable has some percentage of randomly missing data, five variables with small percentages of missing data can add up to a substantial portion of a sample being deleted, which can have deleterious effects on power.
  - Thus, case deletion is only an innocuous practice when:
    - The number of cased with missing data is a small percentage of overall sample, and
    - The data are demonstrably MAR

Preprocessing Pipeline

- Effects of Mean/Median/Mode Substitution
    - In the absence of any other information, the mean is the best single estimate.
    - The flaw in this is that if 20% of a sample is missing, even at random, substituting the identical score for a large proportion of the sample artificially reduces the variance of the variable.
    - And as the percentage of missing data increases, the effects of missing data become more profound

Preprocessing Pipeline

- Imputation
  - Single Imputation
    Assuming most variables have complete data, and they are strongly related to the variable with the missing data, a researcher can create a prediction equation using the variables with complete data, estimating values for the missing cases much more accurately than simple mean substitution.

  - Multiple Imputation
    It uses a variety of advanced techniques for example, maximum likelihood estimation, Markov Chain Monte Carlo (MCM) sampling, etc. These estimate missing values by creating multiple versions of the same data set that explore the scope and effect of the missing data.

Thank you!