AIML213

Assignment Two

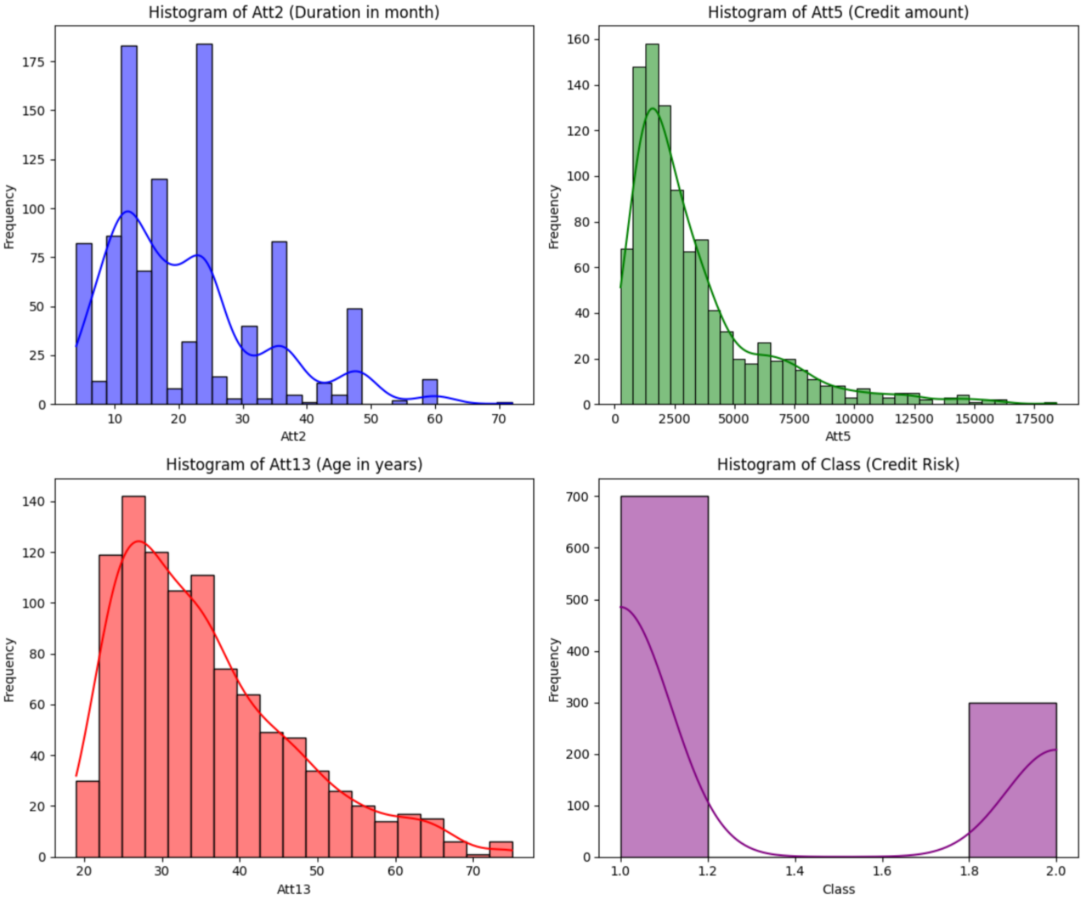
300601546

Shemaiah Rangitaawa
4-26-2024

Understanding the Data

Histograms

When analyzing histograms, I used matplotlib's 'fd' function, which uses the Freedman-Diaconis rule to estimate the proper bin size. This rule is effective for larger datasets because it tailors the bin sizes based on the spread and variability of the data. Specifically, the bin width is decided by the interquartile range and is adjusted inversely with the cube root of the dataset size, ensuring that the bin sizing is best for the given data distribution.



- **Att2** shows a positively skewed distribution with a moderate peak.

- **Att5** is highly positively skewed and has a sharp peak with heavy tails.

- **Att13** also has positive skewness with a slight peak above the normal distribution.

- **Class** displays moderate positive skewness but is flatter than a normal distribution, with lighter tails.

Null Data Entries

The dataset has missing values in one feature, "Att1," totaling 65 entries, which represents 0.31% of the dataset. We will use mean imputation to fill in these null entries during preprocessing.

ML Task

In the context of this task, the algorithm will learn from historical data that includes various attributes of customers, such as their financial behavior, loan history, repayment timelines, and other relevant financial characteristics. These features help the algorithm understand patterns and distinctions between what constitutes good and poor credit risks. The aim is to train a model so that it can accurately predict the credit risk category of new customers based on the learned patterns.

Data Preprocessing

The preprocessing of the dataset was done with several steps to ensure the data's suitability for machine learning models. Initially, ordinal features were found and explicitly ordered to preserve their ranking, a necessary step for proper numerical representation. Subsequently, missing values within these ordinal features were imputed using the mode of each feature, preserving the dataset's consistency while avoiding the introduction of biases.

For categorical features that were nominal, one-hot encoding was applied, with the first category omitted to prevent multicollinearity, which could compromise the model's performance. In contrast, numeric features were addressed by applying median imputation to manage missing values effectively, ensuring the dataset's robustness against outlier influences.

To standardize the numeric data, each feature was transformed to have a zero mean and a unit standard deviation using the StandardScaler from scikit-learn.

Feature Selection

The feature selection process using mutual information showed the following features as relevant to the target output formatted as the feature index with corresponding feature description:

Feature Ranking with SFFS:

33 : Status of existing checking account

34 : Credit history

14 : Housing - own

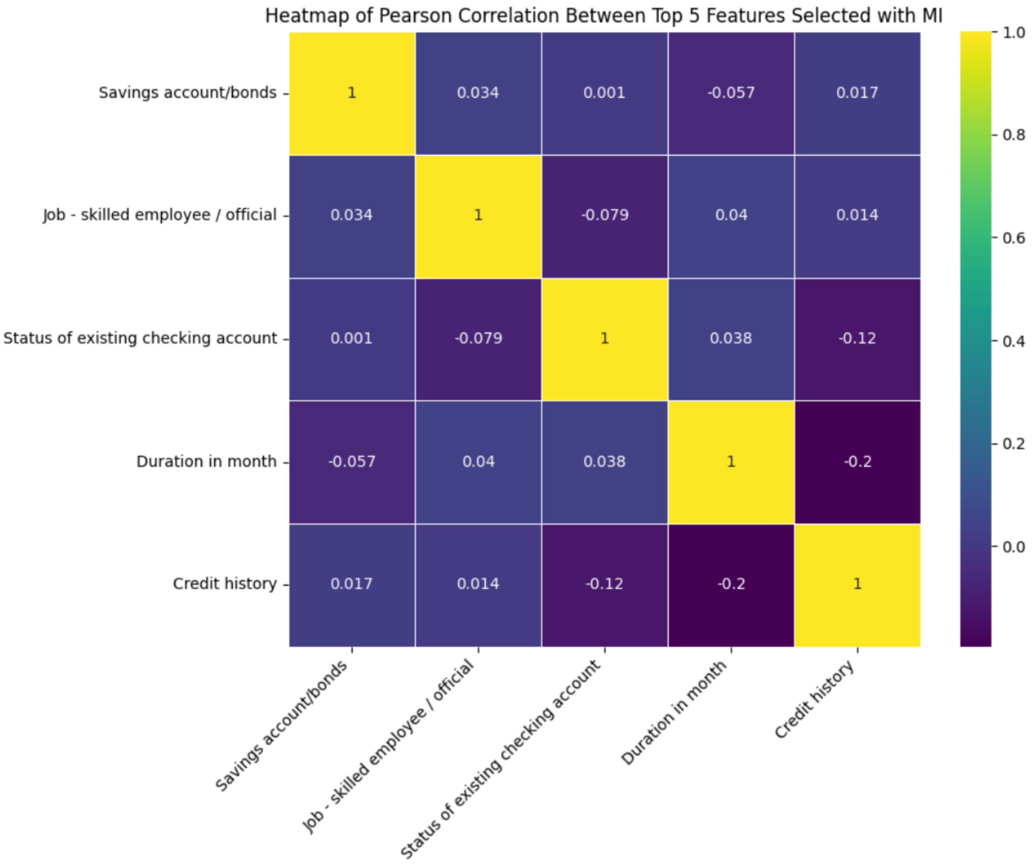
36 : Present employment since

15 : Housing - for free

Accuracy of using all features: 61.12%

Accuracy of using top five features: 69.26%

Using the subset of top five features found through Mutual Information ranking led to an improvement in model accuracy on the test set, from 61.12% with all features to 69.26% with just the top five features. This increase in performance suggests that the selected features provide a more concise yet sufficient representation of the data for the prediction task. It implies that the top features are highly informative and the model benefits from the reduction of noise or irrelevant information that may be present in the less prominent features. Additionally, the reduction in features could also lead to faster computation times and simpler models, which are easier to interpret and less prone to overfitting.



	Savings account/bonds	Job - skilled employee / official	Status of existing checking account	Duration in month	Credit history
Savings account/bonds	1	0.034	0.001	-0.057	0.017
Job - skilled employee / official	0.034	1	-0.079	0.04	0.014
Status of existing checking account	0.001	-0.079	1	0.038	-0.12
Duration in month	-0.057	0.04	0.038	1	-0.2
Credit history	0.017	0.014	-0.12	-0.2	1

The heatmap shows the Pearson correlation between five features selected by Mutual Information, showing how each pair of features is related linearly.

- **Correlations Between Different Features:** The correlations between unique features are all quite low, as most off-diagonal cells are dark purple, standing for values close to zero. This suggests little to no linear relationship between them.
- **Notable Correlations:** The strongest negative correlation is between "Duration in month" and "Credit history" at -0.2, hinting that longer credit durations might be associated with a particular type of credit history. The strongest positive correlation is a very weak one of 0.04 between "Job - skilled employee/official" and "Duration in month", which could suggest a very slight tendency for skilled employees or officials to have longer credit durations.

Sequential Forward Feature Selection

The Sequential Forward Feature Selection (SFFS) algorithm we implemented is a wrapper feature selection approach. The main characteristic of a wrapper method is that it evaluates the predictive power of a subset of features by training a model and computing a performance metric, which is what the sequential_score function does by using cross validation on the training set with the classifier.

In the scoring function, for each subset of features S, a classifier is trained, and its performance is scored on the validation fold. This score is then used to decide which feature to add to the set S.

Wrapper methods are more computationally intensive than filter methods because they require training a new model for each feature subset considered.

In contrast, filter methods evaluate features based on intrinsic properties of the data, such as correlation or mutual information, without involving a predictive model. Embedded methods include feature selection as part of the model training process and make feature selection decisions based on the training process.

Our implementation is a feature subset selection approach, not a feature ranking approach and the distinction lies in the method. In feature ranking, each feature is given a score independently of the others, and then features are ranked based on these scores. The ranking is generally done without considering potential interactions between features.

In feature subset selection, a subset of features is chosen based on the joint predictive power of the features within that subset. It's an iterative process that evaluates the performance of a model with the addition or removal of features in a subset. The goal is to find the best subset that perfects a performance metric, rather than to rank all features.

Feature Ranking with SFFS:

13 : Purpose - education

34 : Credit history

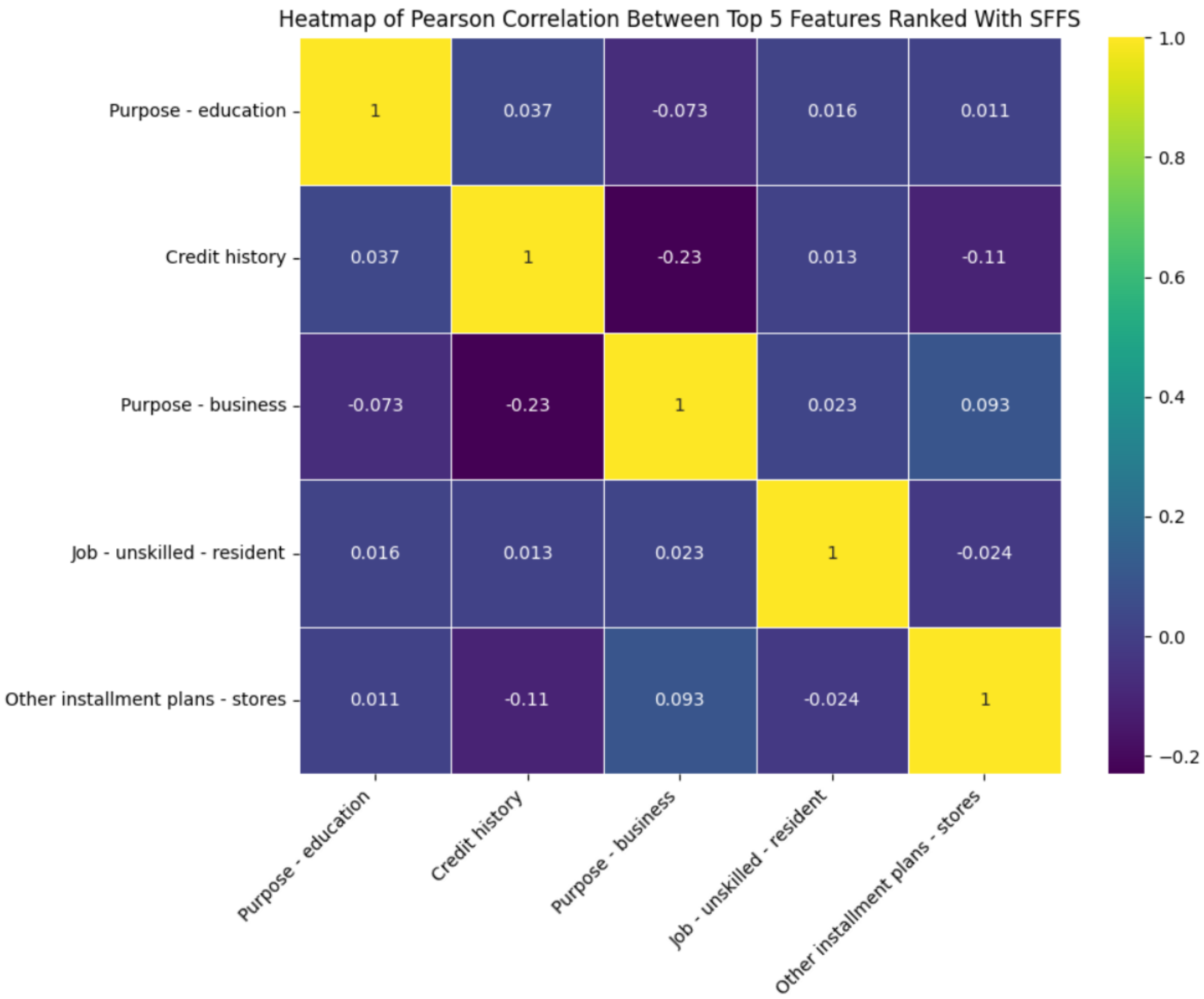
15 : Purpose - business

28 : Job - unskilled - resident

24 : Other installment plans - stores

Accuracy of using all features: 61.12%

Accuracy of using top five features: 52.96%



- **Credit History and Purpose - Business**: The most notable relationship is the negative correlation of -0.23 between "Credit history" and "Purpose - business", which suggests that these features have an inverse linear relationship. When "Credit history" has high values, "Purpose - business" tends to have lower values, and vice versa.
- **Purpose - Business and Other Installment Plans - Stores**: There is a positive correlation of 0.093, showing a small direct linear relationship. Higher values in one feature are slightly associated with higher values in the other.
- **Other Weak Correlations**: The rest of the correlations are relatively weak, close to 0, indicating very little linear relationship between the respective feature pairs.

When comparing the testing performance of the two algorithms, the feature ranking with Mutual Information had a higher accuracy of 69.26% when using the top five features, as opposed to the 52.96% accuracy achieved with the top five features selected by the Sequential Forward Feature Selection (SFFS).

The MI-based approach outperforms SFFS in this case. The higher accuracy shows that the top five features found by MI have a stronger predictive power than those selected by SFFS. This could be due to MI's ability to capture non-linear dependencies between the features and the target variable, which may not be fully considered in a linear model-based approach like SFFS with the classifier used for evaluation.

Based on the accuracy metrics, the feature ranking method using MI is the better approach for this dataset and the classifier used, as it leads to a more accurate model on the test set.