AI351 - Machine Learning Lab

Taiwanese Bankruptcy Statistics

Ali Hasan Khan
Artificial Intelligence(FCSE)
Ghulam Ishaq Khan Institute of Engineering and Technology
Topi,Swabi, Pakistan
u2021079@giki.edu.pk

I. Introduction

In the realm of machine learning and data analysis, one of the paramount challenges often encountered is feature selection and model optimization, particularly in scenarios characterized by a high-dimensional dataset. In this study, we embark on an insightful journey into the world of feature selection and model evaluation, with a focus on the Taiwanese bankruptcy data set—a complex data set comprising 96 columns and 6,800 instances. Our overarching goal is to unravel the intricate interplay between feature selection techniques and classification models, ultimately striving for the identification of the most salient features and the optimal predictive model for this dataset.

To navigate this intricate landscape, we employ a diverse array of methodologies, each meticulously selected to address specific facets of the problem at hand. Our arsenal includes Variance Threshold for feature dimensionality reduction, K-Best feature selection to identify the most discriminative attributes, Logistic Regression as a foundational classification model, Mutual Information Classification for feature ranking, Random Forest Classifier for ensemble learning, Principal Component Analysis (PCA) for dimensionality reduction, and the powerful Sklearn Genetic Algorithm for automated feature selection.

Through systematic experimentation and rigorous analysis, we aim to shed light on the effectiveness of these techniques, with the ultimate objective of providing valuable insights into feature importance and model performance. As we embark on this data-driven odyssey, we invite you to accompany us on this voyage of discovery, where we unravel the intricate tapestry of data, algorithms, and insights, striving to enhance our understanding of feature selection and classification in the context of a real-world dataset.

II. Variance Threshold

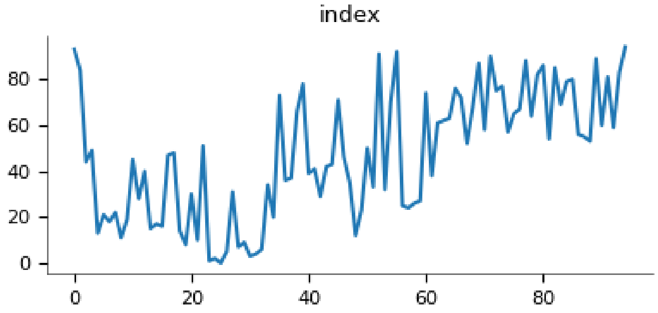
In our pursuit of effective feature selection techniques for the Taiwanese bankruptcy dataset, we turn to the powerful concept of Variance Thresholding. This method, harnessed through Python's scikit-learn library, empowers us to identify and retain features that exhibit substantial variance, while discarding those that remain relatively constant across the dataset.

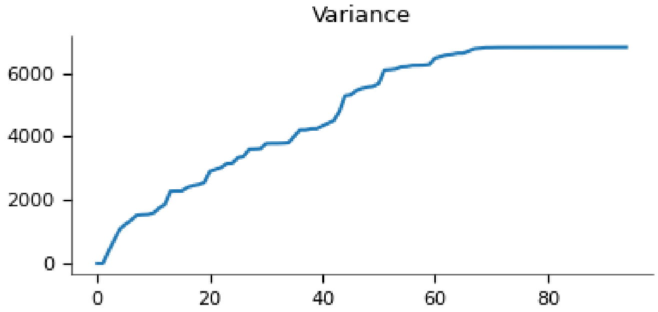
978-1-6654-5992-1/22/$31.00 ©2022 IEEE

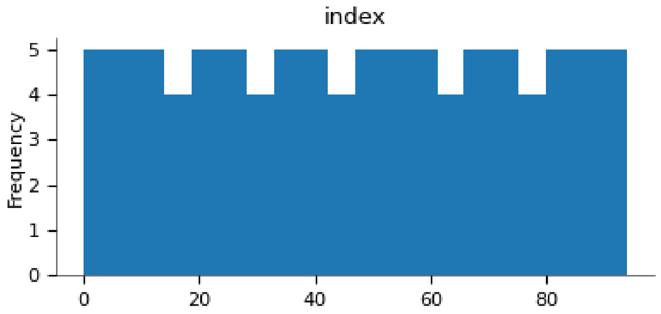
Utilizing a stratified train-test split (with 70 percent allocated to the training set), we proceed to apply the Variance Threshold function. We set the threshold to zero, which is the default value, indicating that any feature with zero variance—i.e., those that remain constant throughout the data-set—will be eliminated. The resulting transformed training data-set, X train vth, contains only the features that exhibit varying values, enhancing the potential for meaningful patterns and predictive power.

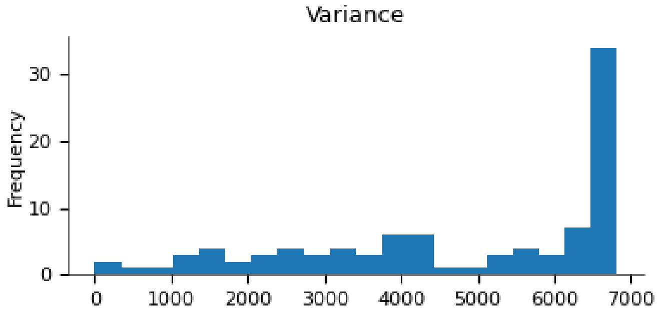
To gain deeper insights into the impact of this process, we present a Data-Frame (df1) showcasing the variance values of each feature after applying Variance Thresholding. This allows us to assess the degree of variance reduction and highlights the features that contribute the most information to the dataset. Through this approach, we aim to identify and retain only the most informative features, streamlining our dataset and potentially improving the efficiency and effectiveness of subsequent classification models.

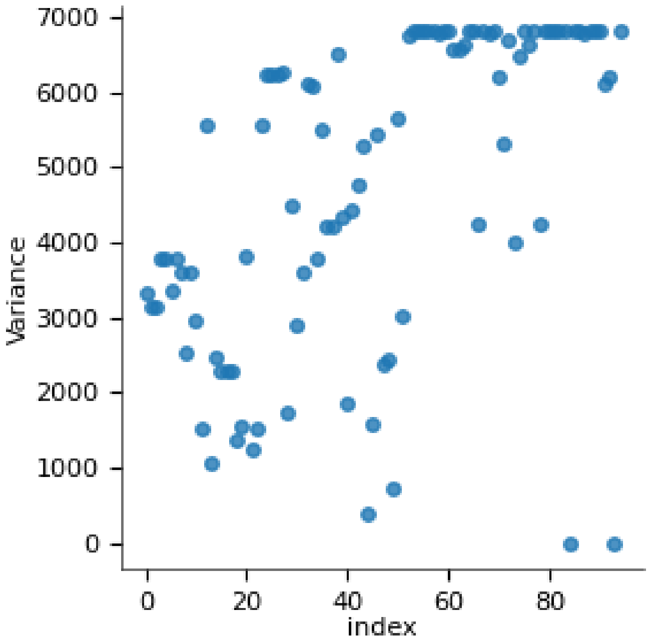
The following figures show the result on the data-set.

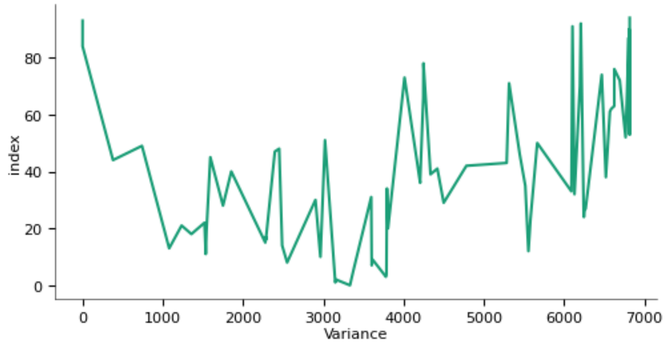


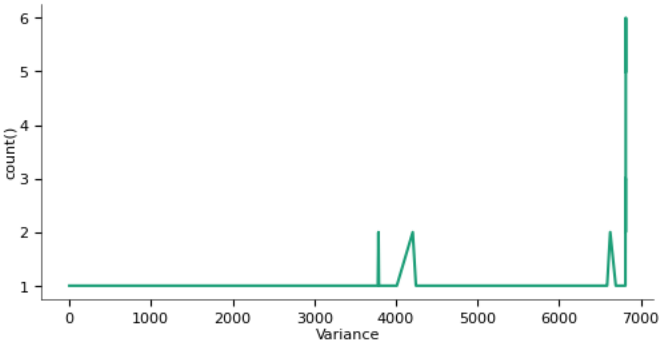








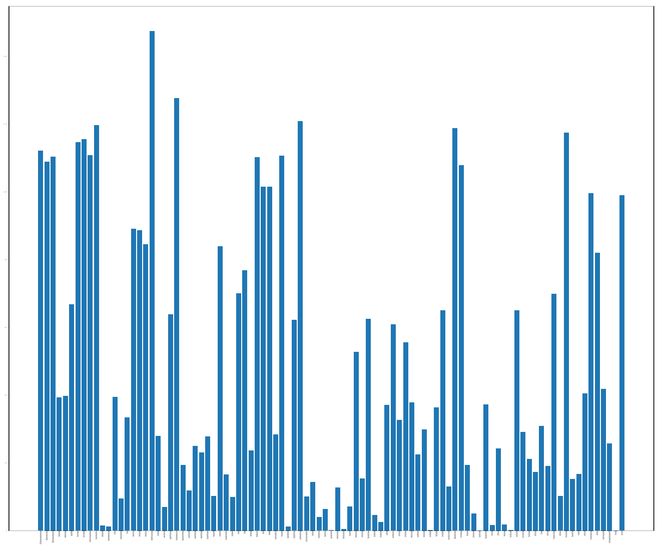




III. SELECTKBEST AND FEATURE SCORING

In our endeavor to pinpoint the most influential features within the Taiwanese bankruptcy dataset, we employ the SelectKBest feature selection method coupled with the F-statistic (f-classif) scoring function. This robust approach, facilitated through scikit-learn, aids in identifying the K most discriminative features for our classification task. In this particular instance, we set K to 6, aiming to retain a concise set of attributes that are most relevant for our predictive model.

As the feature scoring unfolds, we gain valuable insights into the discriminatory power of each feature within our dataset. The F-statistic scores for each feature are meticulously calculated and presented, revealing the extent to which each attribute contributes to our classification problem. Subsequently, a visual representation in the form of a bar chart is generated, providing an at-a-glance overview of feature significance. By scrutinizing these scores, we can make informed decisions about which attributes to include in our final feature subset. This meticulous feature selection process is a pivotal step in enhancing the accuracy and efficiency of our classification models, allowing us to extract the most informative features from the dataset while reducing dimensionality.



In the pursuit of optimizing our feature selection process for the Taiwanese bankruptcy dataset, we employ the SelectKBest method with an F-statistic scoring function (f-classif). Following this rigorous feature selection process, we transform both the training and test datasets to ensure consistency in the features used for subsequent classification tasks. The transformed training set, X-train-classif, reflects the refined selection of features, providing us with a focused subset for training our classification models. We emphasize that this process enhances computational efficiency and potentially improves model performance by retaining only the most relevant attributes. These transformations are essential steps in our data preprocessing pipeline, contributing to more effective and streamlined machine learning workflows.

X-train.shape: (4773, 95)
X-train-selected.shape: (4773, 6)

IV. Logistic Regression and Feature Selection

In our quest to enhance the classification performance on the Taiwanese bankruptcy dataset, we turn to Logistic Regression as a foundational classification model. Leveraging the 'liblinear' solver and setting a random state for reproducibility, we initially apply this classifier to the dataset in its entirety. The result, indicated by the score with all features, provides a baseline performance metric.

However, recognizing the importance of feature selection in improving model efficiency and interpretability, we proceed to deploy the model on the refined dataset with only the selected features—those identified as the most discriminative by the SelectKBest method. This selective approach enables us to assess the impact of feature reduction on classification accuracy.

By comparing the scores between these two scenarios, we gain valuable insights into the role of feature selection in enhancing the predictive power of our Logistic Regression model. This comparative analysis informs us of the trade-offs and benefits associated with feature selection, allowing us to make informed decisions in pursuit of the most effective machine learning solutions.

Score with all features: 0.9633 Score with only selected features: 0.9633

V. Mutual Information for Feature Selection

In our relentless pursuit of identifying the most informative features within the Taiwanese bankruptcy dataset, we employ Mutual Information as a potent tool. Implemented through the SelectKBest function with mutual-info-classif as the scoring function, we seek to discern the intrinsic relationships between attributes and the target variable, with a focus on capturing information gain.

This feature selection process unravels the significance of each attribute by calculating their mutual information scores. These scores are a reflection of how much information an attribute provides about the target variable—higher scores signify greater relevance. To provide a visual representation of these scores, we generate a bar chart, facilitating a comprehensive understanding of feature importance and aiding in informed decision-making regarding feature inclusion or exclusion.

Furthermore, we delve into the SelectPercentile method, which allows us to retain the top percentile (in this case, 50 percent) of attributes with the highest mutual information scores. This strategic approach ensures that we preserve the most salient attributes while reducing dimensionality, a crucial step in enhancing the efficiency and effectiveness of our classification models.

VI. Model-Based and Recursive Feature Selection

Our journey into feature selection techniques continues with a foray into model-based methodologies that harness the predictive power of a RandomForestClassifier. In the pursuit of feature importance, we first employ the SelectFromModel method. This approach allows us to discern feature significance by leveraging a random forest classifier with 100 estimators. Features are selected based on their importance scores, specifically targeting those surpassing the median threshold.

Moving forward, we explore the realm of Recursive Feature Elimination (RFE) with yet another RandomForestClassifier. With a goal of retaining the most influential attributes, RFE iteratively evaluates feature importance and progressively eliminates the least informative features until the desired number is achieved.

Lastly, we delve into the power of Recursive Feature Elimination with Cross-Validation (RFECV). In this approach, we embrace a RandomForestClassifier and subject it to 5-fold cross-validation. RFECV performs a meticulous feature ranking and selection process, ensuring that the optimal subset of attributes is chosen based on their contribution to model accuracy.

Through these advanced techniques, we aim to distill the essence of feature importance, shedding light on which attributes hold the most predictive potential for our classification task. This exhaustive exploration equips us with valuable insights into feature selection, enabling us to refine our dataset and enhance the efficiency and effectiveness of our machine learning models.

Optimal number of features : 14

VII. Correlation Analysis

To gain deeper insights into the relationships between different financial indicators and the likelihood of bankruptcy, we turn to correlation analysis. Using the Pandas library, we create a DataFrame, df-corr, to store the correlations between each attribute and the 'Bankrupt?' target variable. This allows us to assess the strength and direction of these relationships.

Visualizing the correlations is paramount to our analysis. Employing a color-coded bar chart, we highlight the attributes with positive correlations in light green, providing a clear visual distinction. Additionally, we sort the correlations in ascending order to emphasize the attributes with the most negative correlations. This graphical representation not only aids in identifying potentially influential financial indicators but also lays the groundwork for understanding their impact on the bankruptcy prediction task.

Correlation analysis serves as a pivotal step in feature selection and model development, as it informs us of the attributes that exhibit the strongest statistical relationships with the target variable. These insights guide our decision-making process when selecting the most relevant features for our machine learning models, thereby contributing to the overall effectiveness of our predictive models.

VIII. Mutual Information for Feature Selection

Our quest for identifying the most informative features in the Taiwanese bankruptcy dataset continues, this time through the lens of Mutual Information. Employing scikit-learn's mutual-info-classif, we calculate the mutual information scores for each attribute in relation to the 'Bankrupt?' target variable. These scores are a reflection of the amount of information that each feature holds about the target, thereby revealing their significance in our classification task.

To facilitate a comprehensive understanding of these scores, we create a DataFrame that pairs each attribute's name with its corresponding mutual information score. This structured representation allows us to discern which features exhibit the highest levels of information gain in predicting bankruptcy.

Additionally, we transform this numerical information into a visual format by plotting the mutual information scores. This graphical representation provides a clear overview of attribute importance, enabling us to make informed decisions about which features to prioritize in our machine learning models.

Mutual Information serves as a valuable tool in our feature selection arsenal, guiding us toward a refined subset of attributes that possess the greatest predictive power. This meticulous process enhances the efficiency and accuracy of our classification models and plays a pivotal role in shaping the success of our predictive analytics endeavors.

IX. Genetic Algorithm for Feature Selection

In our pursuit of optimal feature selection for classification tasks, we employ a powerful technique known as Genetic Algorithm-based feature selection. This code snippet showcases the application of GeneticSelectionCV, a module that leverages the Genetic Algorithm to iteratively evaluate subsets of features for a given classification model.

To begin, we load our dataset, the Taiwanese bankruptcy dataset, and preprocess it to include the first 59 columns as attributes. Additionally, we introduce some random noisy data to the dataset, adding complexity to the feature selection process.

For our classification model, we utilize logistic regression with multi-class support. The heart of our feature selection process lies in the GeneticSelectionCV module. Here, we configure several parameters, including population size, mutation probabilities, and generations, to fine-tune the algorithm's behavior.

Upon execution, GeneticSelectionCV evaluates subsets of features and aims to maximize the classification accuracy through genetic operations like crossover and mutation. The resulting selector.support attribute provides a binary mask indicating which features are selected for the final model.

This Genetic Algorithm-based approach offers a systematic and efficient method for optimizing feature selection, especially in scenarios with high-dimensional datasets. By embracing this technique, we seek to identify the most relevant attributes and enhance the predictive power of our classification models.

X. Dimensionality Reduction with PCA

In our journey towards extracting meaningful insights from high-dimensional data, we harness the power of Principal Component Analysis (PCA). This code snippet demonstrates the application of PCA as a dimensionality reduction technique to streamline our dataset and gain a clearer understanding of the underlying variance.

To prepare our data for PCA, we first standardize it using the StandardScaler from scikit-learn. Standardization ensures that all features have a mean of zero and a standard deviation of one, creating a level playing field for PCA.

Next, we specify the number of principal components we want to retain—in this case, five (num-components=5). PCA then takes center stage, fitting to the scaled data and transforming it into a new reduced-dimensional space.

One of the most valuable insights from PCA is the explained variance ratios, which we obtain and print. These ratios shed light on the proportion of total variance accounted for by each principal component. In essence, they allow us to discern the significance of each component in explaining the dataset's overall variability.

PCA serves as a crucial tool in feature reduction and visualization, enabling us to condense complex data into a more manageable form while retaining the most informative aspects. This reduction in dimensionality can improve computational efficiency and help identify the most salient patterns and trends within the dataset, furthering our data exploration and modeling endeavors.