APS Failure Prediction

Berat Dalsuna
Computer Engineering
Istanbul Technical University
Istanbul, Turkey
Email: dalsuna20@itu.edu.tr

Semih Gençten
Computer Engineering
Istanbul Technical University
Istanbul, Turkey
Email: gencten20@itu.edu.tr

Yasin İbiş
Computer Engineering
Istanbul Technical University
Istanbu1, Turkey
Email: ibisy20@itu.edu.tr

***Abstract*—For the APS Failure Prediction dataset, dimensionality reduction techniques are implemented, features are eliminated or generated, oversampling is applied, logistic regression and random forest models are analysed.**

I. Introduction

In this paper, APS Failure Prediction dataset is analysed and best prediction is tried to be achieved with several models and methods. Dimensionality reduction methods, data exploration, feature engineering and model selection are important aspects of this study.

II. Data Exploration

A. Exploratory Data Analysis

When analysing the data, correlations between the feature and target variables, outliers, balancing of the classes and null value percentage for each feature are considered.

Firstly, value counts for the classes are checked. There are two classes called 'pos' and 'neg' in this dataset. It is observed that while there are 59000 instances for the negative class, there exists 1000 instances for the positive class which results in very high imbalance between the positive and negative classes.

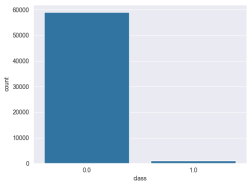

Fig. 1. Class distribution

Secondly, null value percentages checked for each feature. For first fifty features with the most percentages, values are plotted with bar plot. It can be seen below in figure 2.

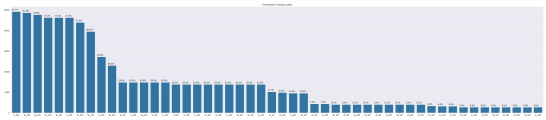

Fig. 2. Missing feature counts

From the above plot we see,

- 8 features have missing values greater than 60
- 16 features have missing values between 20
- the remaining features have missing values less than 20

Thirdly, outliers are detected with two models which are called LocalOutlierFactor and IsolationForest. Since it is not a normal distributed dataset, trivial methods such as quartile range are failed to detect outliers. Before fitting the data for the models, median imputing strategy is used to fill null values.

For the LocalOutlierFactor, nearest neighbors are checked for every data. If it is too far from the neighbors, it is flagged as outlier. 3170 outliers found with this method. For the isolation forest algorithm, points are isolated by random splits and random thresholds. 2765 outliers found with this method. Amount of the intersecting outliers for mentioned methods are 291.

Lastly, feature and target correlations are analysed. For every feature, correlation with the class column is calculated. If it is negative or positive, it can be said they are positively correlated or negatively correlated. But if it is near to zero, they are uncorrelated. Top 5 features for this correlation are:

1) ci_000 with value 0.553
2) bb_000 with value 0.542
3) bv_000 with value 0.541
4) bu_000 with value 0.541
5) cq_000 with value 0.541

Plots for the correlations in a descending order:

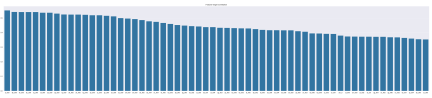

Fig. 3. First 50 target feature correlations

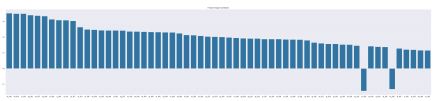

Fig. 4. Middle target feature correlations

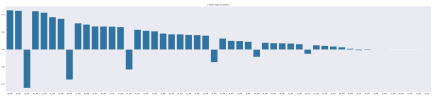

Fig. 5. Last target feature correlations

B. Conclusion

Summarizing the conclusions drawn from Exploratory Data Analysis:

- Data is highly imbalanced with almost 98 percent of the data points belonging to the negative class.
- There lots of mising information in data which makes harder to work with.

III. Methodology

A. Preprocessing

In preprocessing step, 3 different dimensionality reduction techniques (PCA, LDA, Factor Analysis) are implemented and analysed.

1) PCA: Principal component analysis summarizes the information, trying to hold maximum variance with decreased amount of components with the help of eigenvalues. Number of components selection is hard to make. Explained variance is a good measure for deciding how many components to choose. Explained variance plot can be seen below:

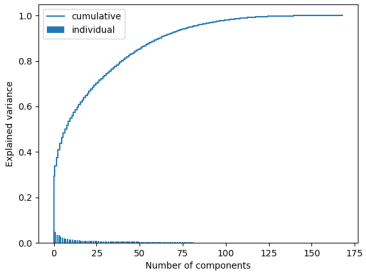

Fig. 6. Explained variance

Two and three dimensional plots can be visualized. Different colors are used for positive and negative classes.

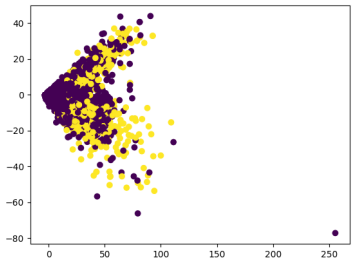

Fig. 7. 2-D PCA

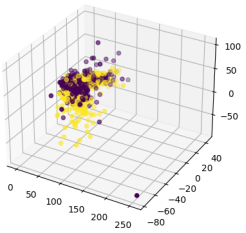

Fig. 8. 3-D PCA

A correlation analysis is done for the PCA. Scatter Coefficient is calculated and determinant checked for the covariance matrix. It is equal to 1.702817572377432e-173 which is a very low value. It indicates PCA can be useful but LDA performed better, so it is selected for dimensionality reduction.

2) LDA: Linear Discriminant Analysis is a dimensionality reduction technique which is used in supervised learning. LDA aims to find a linear combination of features that characterizes or separates two or more classes in the data. The method achieve this by maximizing the spread between classes and minimizing the spread within each class.

Linear Discriminant Analysis (LDA) accomplishes dimensionality reduction with a focus on class separation. It begins by computing within-class scatter (S_w) and between-class scatter (S_b) matrices. By maximizing the ratio of the determinant of S_b to S_w, LDA identifies linear combinations of features that maximize the distance between class means and minimize within-class spread. The method, through eigendecomposition of $S_w^{-1} \cdot S_b$, selects the top eigenvectors to form a reduced-dimensional space. Projecting data onto this space retains discriminatory information, ensuring that the transformed features are optimized for effective classification.

Since it seperates the classes it can be seen from the below graph, it is a better preprocessing technique than PCA.

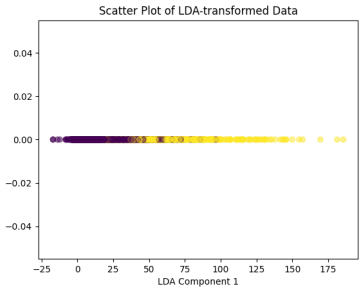

Fig. 9. 2-D LDA

3) Factor Analysis: Factor Analysis is a statistical method used to identify underlying latent factors influencing observed variables. It assumes that the observed variables are linear combinations of unobservable factors and error terms. The goal is to uncover these latent factors and understand their

impact on the observed data. Factor Analysis involves decomposing the covariance or correlation matrix of observed variables into factors, factor loadings, and unique factors. By performing eigenvalue decomposition and possibly rotating the factor loadings, the method seeks a more interpretable representation of the data. High factor loadings indicate strong associations between observed variables and latent factors. Factor Analysis helps reduce the dimensionality of the data, revealing hidden structures and simplifying the interpretation of complex datasets, making it valuable in fields such as psychology and social sciences.

We applied a test to decide wheter to use factor analysis or not. The KMO test helps us figure out if our data is good for factor analysis. We got a KMO value of 0.53, which Kaiser says is ”miserable.” This means our variables don't have much in common, and it might not work well for factor analysis. Having a KMO below 0.6 is generally not good for this analysis. So, we need to be careful before using factor analysis with this data. It's a good idea to check our variables again, look at how we collected the data, or maybe use a different analysis method that fits our data better.

According to Kaiser, a KMO below 0.6 is generally considered insufficient for robust factor analysis. The obtained result of 0.53 implies weak correlations or limited shared variance among variables, challenging the assumptions fundamental to factor analysis.

Therefore we decided to not use factor analysis as a preprocessing technique.

B. Feature Engineering

Feature engineering includes removing unnecessary data, imputing missing values and oversampling.

1) Removing Data: Duplicate rows are removed from dataset. Also columns with zero variance which means having same value for every row are deleted.

2) Imputing Data: For imputing, two strategies (mean, median) are used. For mean strategy, null values are filled with the mean of the column. For median strategy, they are filled with the median value. Columns are selected by looking at the percentage of the null values for each column.

The imputation strategy that we follow is:

- We will eliminate features with missing values greater than 60
- We will perform median imputation of features with missing values less than 40

3) Oversampling: Oversampling is a method to balance dataset by increasing the number of the class which has less samples. For this dataset, number of positive samples are very lower than number of negative samples. After oversampling, number of positive samples is increased to 29500 and number of negative samples decreased to 49166. With this method, models can learn better from the imbalanced dataset.

C. Model Building

Different models are analysed with preprocessed data. Models built with logistic regression, support vector machine and random forest algorithm.

1) Logistic Regression: Stochastic gradient descent is used for optimizing the model. Alpha value set as 0.1 and for loss function log loss is selected. As a post-processing method, sigmoid function is chosen.

2) Support Vector Machine: All hyper parameters are same as logistic regression except the loss function. For the loss function, hinge loss is selected.

3) Random Forest: For random forest, best hyper parameters which includes maximum depth, number of estimators and minimum samples to split are found with a randomized search algorithm. After hyper parameters found, classifier is calibrated using sigmoid function.

IV. Results

For analysing the results, confusion matrix and f1-score is used. f1-score is useful when dealing with imbalanced datasets since it considers both precision and recall.

- For the logistic regression, f1 score is equal to 0.93.
- For the support vector machine, f1 score is equal to 0.94.
- For the random forest, f1 score is equal to 0.89.

Corresponding confusion matrices can be seen below.

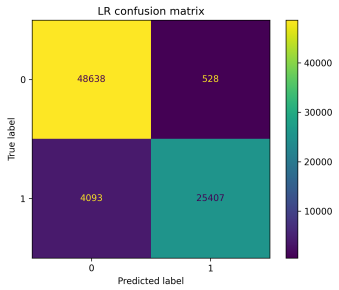

Fig. 10. LR confusion matrix

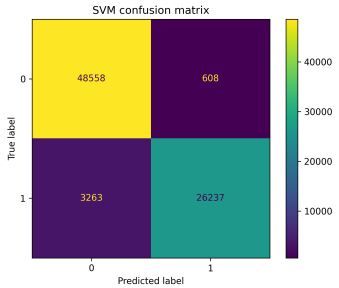

Fig. 11. SVM confusion matrix

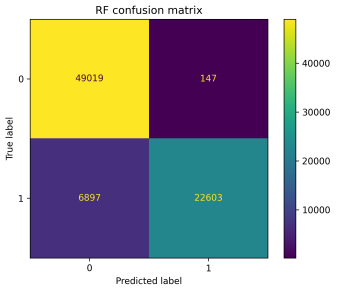

Fig. 12. RF confusion matrix

From above metrics, SVM and LR performs nearly same and have a good scores. So, LR or SVM can be tested on the testing dataset for predictions.

V. Conclusion

This document provides a comprehensive exploration of the APS Failure Prediction dataset, emphasizing data preprocessing, feature engineering, and model building. Notable findings include a highly imbalanced class distribution, the identification of outliers using various algorithms, and insightful visualizations of feature correlations. The methodology incorporates techniques like PCA, LDA, and Factor Analysis, with a thorough analysis of their effectiveness. This paper serves as a valuable resource fhandling imbalanced datasets.

References

[1] Dash, S. K. (2023, November 20). A brief introduction to linear discriminant analysis. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/

[2] Sklearn.decomposition.PCA. (n.d.). Scikit-learn. https://scikit learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[3] Jaadi, Z. (2023, March 29). A Step-by-Step Explanation of Principal Component Analysis (PCA). Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[4] Statistics Solutions. (2021, August 10). Factor Analysis - Statistics Solutions. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis

[5] Patil, P. (2022, May 30). What is Exploratory Data Analysis? - Towards Data Science. Medium. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15