BLG454E Term Project: APS Failure Prediction

Due: 01.01.2024 23.00

1 Introduction

Esteemed participants, welcome to the "APS Failure Prediction" term project—an immersive foray into the realm of applied machine learning. In this undertaking, you will navigate the complexities of the IDA2016Challenge dataset. You will participate in a Kaggle challenge from this link.

1.1 Project Overview

Your mission is to discern the fate of the Air Pressure System (APS) within heavy Scania trucks. The dataset, drawn from the daily operations of these vehicles, presents a binary classification challenge. The **positive class** signifies component failures within the APS, while the **negative class** represents failures unrelated to this critical system.

1.2 Challenge Metric

However, we aren't merely concerned with accuracy. We introduce a cost metric of average precision to infuse practicality into your predictive models.

1.3 Dataset Information

In your arsenal, you possess a training set comprising 60,000 examples—59,000 negatives and 1,000 positives. The test set awaits your predictions with 16,000 examples. Boasting 171 attributes, including 7 histogram variables, this dataset is a fertile ground for your prowess in feature engineering and model refinement.

1.4 Project Components

Now, let us delineate the components that constitute this intellectual odyssey:

1. **Kaggle Challenge (60 points + Bonus):** Your primary objective is to implement a machine learning model for APS Failure Prediction. Subsequently, submit your predictions to Kaggle. Bonus points await the top performers. Will you rise to the challenge?

2. **Report Writing (40 points):** The Latex report serves as your narrative canvas. Document your journey comprehensively—data exploration, methodology, results, and conclusions for both the Kaggle challenge and the additional coding problem.

Prepare yourselves for an intellectually invigorating expedition. This project transcends the mere construction of models; it necessitates judicious decision-making and effective communication of your insights.

2 Kaggle Challenge (60 points + Bonus)

2.1 Subproblem: Custom Feature Scaling and Preprocessing from Scratch (20 points)

In this preparatory subproblem, you are tasked with implementing a custom feature scaling method without relying on external libraries. The objective is to create a scaled version of the dataset that will be used as input for the Kaggle challenge. This exercise aims to deepen your understanding of feature scaling and enhance your coding proficiency.

2.1.1 Subproblem Definition

The subproblem revolves around the implementation of a custom feature scaling method. Feature scaling is a crucial preprocessing step that standardizes the range of independent variables. In this context, you are required to develop a custom scaling algorithm tailored to the characteristics of the APS Failure dataset, apply 3 different dimensionality reduction techniques (PCA, LDA, Factor Analysis) and decide whether to proceed with your reduced dimensions or not. Then you will divide your dataset to folds to apply cross-validation.

2.1.2 Implementation Guidelines

To successfully tackle this subproblem, adhere to the following guidelines:

- **Code from Scratch:** Implement the custom feature scaling and dimensionality reduction methods without using external libraries. Develop the preprocessing from scratch, showcasing your coding skills.
- **Algorithmic Clarity:** Clearly document the algorithmic steps of your implementation. Provide comments and explanations to enhance readability.
- **Correct Expression Considerations:** While external libraries are off-limits, discuss the benefits and drawbacks of using dimensionality reduction techniques you applied based on their level of representation of the actual dataset. Decide on whether to proceed with the reduced dimensions and justify your decision.
- **Integration with Kaggle Challenge:** Demonstrate how the scaled dataset resulting from your custom feature scaling method will be seamlessly integrated with the main machine learning model developed for the Kaggle challenge.

Successful completion of this preparatory subproblem is essential, as the scaled dataset will serve as input for the main Kaggle challenge. Your ability to implement critical preprocessing steps without external libraries will be integral to the success of the overall project.

2.2 Building a Machine Learning Model (40 points)

Your journey continues with the implementation of a robust machine learning model for predicting APS failures. Leverage your skills in data preprocessing, exploratory data analysis, and feature engineering to unravel the intricacies of the IDA2016Challenge dataset. Your goal is to distinguish the **positive class**, representing APS component failures, from the **negative class**, indicative of failures unrelated to the APS.

Implementation Guidelines

You are allowed to use external libraries of your choice in this part. To guide your efforts, consider the following steps:

- **Data Preprocessing:** Use the preprocessed data you obtained after completing the previous subproblem.

- **Exploratory Data Analysis (EDA):** Delve into the dataset's nuances through EDA. Visualize key features and relationships by correlation analysis, outlier detection and feature vs. target relationships to gain insights into potential patterns.
- **Feature Engineering:** Elevate your model's predictive power through thoughtful feature engineering. Consider transforming existing features and creating new ones, improving your dimensionality reduction analysis.
- **Model Building:** Implement a machine learning model of your choice. Optimize hyperparameters and evaluate its performance using appropriate metrics.
- **Kaggle Submission:** Submit your predictions to Kaggle. Aim for accuracy, but keep in mind the cost-metric implications.

2.3 Bonus Points (20 points)

For those seeking an extra challenge, bonus points await the top 5 teams in the Kaggle competition. Stand out from the crowd, and you shall be duly rewarded as: 20, 16, 12, 8, 4 extra points for top 5 teams.

Prepare to showcase your analytical acumen and coding prowess. The journey has just begun!

2.4 Evaluation Criteria

Your submission for these problems will be evaluated based on the following criteria:

- **Correctness:** Ensure that your methods provide a reasonably good performance.
- **Algorithmic Clarity:** Evaluate the clarity of your code and the accompanying explanations. A well-documented implementation is crucial. Make sure to add comments to your code wherever needed.
- **Efficiency:** Consider the efficiency of your implementation, especially given the constraints of not using external libraries for the first subproblem.
- **Integration:** Clearly demonstrate how the scaled dataset resulting from your custom feature scaling method will be seamlessly integrated with the main machine learning model developed for the Kaggle challenge.

3 Report Writing (40 points)

The Latex report is your opportunity to showcase the depth and breadth of your work. It serves as a comprehensive documentation of your journey through both the Kaggle challenge and the additional coding problem.

3.1 Report Structure

Ensure your report includes the following sections:

1. **Introduction:** Briefly introduce the problem statement, dataset, and the objectives of your analysis.
2. **Data Exploration:** Provide insights gained from exploratory data analysis. Highlight key visualizations and patterns observed in the data in Kaggle challenge part.
3. **Methodology:** Describe the approach taken in both preprocessing step and the Kaggle challenge. Explain the rationale behind your choices, including algorithms, hyperparameters, and any optimizations.
4. **Results:** Present the results of your machine learning models. Include metrics, visualizations, and any noteworthy observations.
5. **Conclusion:** Summarize your findings and draw conclusions. Reflect on the challenges encountered and lessons learned.

3.2 Formatting Guidelines

For your reports, use the following IEEE conference templates. Ensure that your report is well-structured, clear, and concise. Use appropriate Latex formatting for equations, figures, and tables. Your ability to communicate complex ideas effectively is a key aspect of this evaluation. Prepare to convey your insights in a compelling and articulate manner.

4 Submission

You will submit your code to Kaggle for evaluation. Additionally you will submit your code file together with your report as a zip file. The name of the zip file must contain the Student No of all members seperated by "_", e.g. "150200000_150200001.zip".