Group 7 SHAO Zhihao

Summary

The project is about the dataset - home credit default risk on Kaggle. The objective of the project is to predict repayment abilities. He proposed a clear problem in machine learning, to use a full set of features without selection or a small subset after selection. He employed several statistical and machine learning approaches to examine this problem. The main training dataset is used(application_train). Several important statistical assumptions are checked to ensure the validation of the model, including normality, multicollinearity, missing values, and correlations. As for the feature engineering part, PCA analysis is conducted and 16 most important features are selected. Then, three major models are adopted on the full predictors and the 16 most important features, which are logistic regression, linear discriminant analysis(LDA), and random forest. The result is interesting. where the "premium" subset of features performs better than the full set of features in logistic regression while having similar results in LDA or random forest. LDA with all predictors is selected to be the optimal model with 0.75 AUC on the validation dataset.

Strength of the report

Strong analytical skills are demonstrated as the data exploration, modeling techniques, and feature selections are useful for figuring out the problem statement. Readers can easily understand the problems and solutions by following the report in each subsection. Insightful analysis is also given after the model results.

Weakness of the report

Model hyperparameters and the "premium" subset of features are not given, making it difficult for readers to reproduce the results and continue the project. Motivation behinds the selection of certain models may be discussed more, as readers may be interested in the advantages or disadvantages of selecting a certain model.

• Evaluation on quality of writing (1-5): Is the report written? Is there good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.

5

The report is written and well-organized. Multiple useful figures and graphs are included to explain the result, including the correlation heatmap, ROC curve for demonstrating results, etc. The language is accurate with no grammar mistakes found. I would like to suggest putting more focus on the exploratory data analysis instead of which datasets are chosen since readers are more interested in the statistical assumptions of the data.

• Evaluation on presentation (1-5): Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.

5

The presentation is clear and easy to follow. The language is precise and highly persuasive. Slides are well-organized with a clear table of contents. Overall, the presentation is clear and can be adopted as a tutorial to the newbie of machine learning. I would like to suggest that listing those 16 features and possibly providing reasons for they explain the ability to the response variables.

• Evaluation on creativity (1-5): Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer, you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations, or extensions of existing ideas are also valuable.

5

This work proposed an interesting problem statement that whether a premium subset of features can outperform the full set of features. This is a hot debate in the data science community as it can significantly reduce the effort of collecting more data points for more features. Therefore, the analytical approach and the conclusion given in this project are creative and valuable to the data science community.

• Confidence in your assessment
3, I have carefully read the paper and checked the results