Group 3 review

Summary:

The paper discusses approaches taken in solving the Home Credit Default Challenge. The authors conducted exploratory data analysis and used new correlation coefficient to analyze how categorical features are related. The correlations are later used for feature selection, a LightGBM model is then trained with this subset of features. The final submission scored 0.758 AUC on Kaggle.

Strengths:

The paper correctly identifies that using Pearson correlation between categorical features is ill-founded and proceeded to explore new methods to calculate correlation. It is also interesting to see the usage of Bayesian optimization for parameter tuning. Overall, we can see that the authors are keen to explore new techniques in their endeavors, which is appreciated.

Things to improve:

In their correlation analysis, the authors used ϕ_k coefficient to measure correlation between categorical variables, they did not define ϕ_k nor did they explain how readers can interpret the values of the coefficients. ϕ_k is the enables non-linear correlation analysis as claimed by the authors, yet we do not see supporting evidence for such a claim, it is also peculiar that they did not apply ϕ_k on numerical features.

As the authors noted in the paper, missing value is a critical issue in this challenge. Unfortunately, it is unclear what the authors did to address this issue during model training or inference.

Many typos and figure mislabeling are present in the paper, which makes reading slightly more difficult.

Clarity: 3

Technical quality: 3

Overall rating: 3

Confidence: 3