MSBD 5013 Statistical Prediction - Project 1 Rebuttal

Group Members:BU Shihan , Li Yilin , Yang Yuxin

For reviewer 1 Cen Xinxin,

- For a). Typo corrected in the updated report. Thank you.
- For b). It's a good point that LightGBM can handle categorical features directly.

However, we did not have enough time to test out if one-hot coding it will reduce the performance or not given the tight timelines. If we have more time we would love to test it out.

For reviewer 2 Chengpakhei,

- H For the first point, indeed we did not go into details about the calculations of ϕ_k. Thank you for pointing this out. We have updated our report to include the reference paper Baak, M., et al. "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics." *Computational Statistics & Data Analysis* 152 (2020): 107043. And the process to calculate ϕ_k is as follows,

Procedure description 1: the calculation of ϕ_K

1. In case of unbinned interval variables, apply a binning to each one. A reasonable binning is generally use-case specific, and needs to be chosen such that the bin width is small enough to capture the variations observed in the data. As a default setting we take 10 uniform bins per variable.
2. Fill the contingency table for a chosen variable pair, which contains N records, r rows and k columns.
3. Evaluate the χ^2 contingency test using the Pearson's χ^2 test statistic (Eq. (7)) and the statistically dependent frequency estimates, as detailed in Section 3.1.
4. Interpret the χ^2 value as coming from a bivariate normal distribution without statistical fluctuations, using Eq. (17).
   i. If $\chi^2 < \chi^2_{\text{ped}}$, set ϕ_K to zero.
   ii. Else, with fixed N, r, k, invert the $X^2_{\text{b.n.}}$ function, *e.g.* using Brent's method (Brent, 1973), and numerically solve for ρ in the range $[0, 1]$.
   iii. The solution for ρ defines the correlation coefficient ϕ_K.

For the second point, we did address the missing value inference in section 2.2 Data Cleaning/ Missing Value Filling.

For the third point, we have proof read the report and corrected some of the typos.

For reviewer 3 Chen Tian,

- For the question 1, for the feature enrichment, about the 8 numerical features. Firstly, we have not add these 8 numerical features and come out with the about 72% accuracy. Then we tried to find is there any method to improve such accuracy. We reviewed the lecture mafs6010z AIFin taught by Yao professor before. One project in that course used some financial domain knowledge to enrich the features and used these eight features which improved the result efficiently. Then we followed them and chose these eight features. In the final conclusion we emphasized that domain knowledge sometimes is important to the final result.
- For the question 2, It's a good question. We have explained in the report part 2.2.3 feature encoding. Since categorical features cannot be directly processed, we need to encode them so that they can be fed into our model. We divided categorical features into two categories: Bi_classes, Multi-classes we encode them in different methods which we have written in our report and code. After encoding, the dimension of the features get larger. From 27 to be 125. Thanks!

For reviewer 4 Liu Yi,

- For the recommendation that we need to state the strategy clearly in the feature selection and construction part. We have done this part of jobs in the code part. We have used a lot of visualization methods to show the process of selecting features. For example, using p-value of Chi-square test to judge the importance of each categorical feature and correlation heatmap for numerical feature and even talked about the overlap of different tables. Maybe next time we will write such part into the report. Thanks for your advice.

For reviewer 5 Yuenzhikun,

- For the 1st point of weakness, we use the correlation matrix heatmap for co-linearity detection among features which we indicated in our report and we also explained in our code notebook. We used these co-linearity characteristics for keeping only one feature if multiple features have high co-linearity, which is one of the useful methods for feature engineering in field of data science. It is true that less features will be more visually accurate for visualization, but the features are indeed of a large amount and are needed to be explored at first glance in EDA part for our feature selection, which we think it is necessary.
- For the 2nd point, we indeed processed the missing values. As indicated by our exploratory visualizations, we deleted features with high percentage of missing values, we also filled the missing values of those high correlations features with mean value within features, not just easily delete them.
- For the 3rd point, we do not think it is a weakness to talk about the LGBM method to almost one page. First, we are completing a report, not a poster. For a report, it is essential to illustrate our method in detailed. Besides, one page includes our visual channel for explaining the whole process of LGBM, which we think is also a good way for directly illustration for reader since we cannot sure that every reader of our report is familiar with LGBM method.

  Thank you!