Proposal - STAT3612 Statistical Machine Learning

Group Members:

- Leo Liu Yuxiang - 3035532894
- Nurdaulet Kemel - 3035596307
- Saalim Mohamed Abdulla - 3035445364
- Wu Zijing - 3035556644
- Naina Srivastava - 3035453414

Given the anonymized Home Equity Line of Credit (HELOC) loans dataset, together with 23 raw features, we've decided to split the project into 5 phases:

1. **Intuitive understanding:**
   Plot out the covariance, correlation, distribution of the entire dataset, including inter-variables' correlation, to get a general sense of how the dataset is like. This could also serve as a later reference after we've performed dataset exploration and some model fitting, to see if the results we've acquired fits this intuitive derivation, for better model explainability.

2. **Data Pre-processing:**
   Depending on the specific result we've acquired in the Intuitive understanding part, if several variables have high correlation, we need to specially process them. Overall speaking, for better runtime performance, we'll take one of the following methods to either achieve dimensionality reduction or feature extraction:
   - PCA (Principal Component Analysis)
   - Random Forest
   - Sequential Forward Selection
   - Sequential Backward Selection
   - LLE (Locally Linear Embedding)

3. **Model selection and fitting:**
   After preprocessing the data, based on the monotonicity requirement, we'll have the following models to fit the data, and try to train a model that can achieve as high of an accuracy score as possible out of these models.
   1. With Monotonicity constraints:
      a. Random Forest (With XGBoosting or LightGBM)
      b. GAM (B-Spline with Monotonicity constraint)
   2. Without Monotonicity constraints:
      a. Random Forest
      b. KNN (K-Nearest-Neighbors)
      c. SVM
      d. Multiple Linear Regression (GAM)

4. **Evaluation and Tuning:**

This will be a repeated process for us to test out different hyperparameters and model selections/combinations to achieve the best possible prediction accuracy. Several evaluation metrics listed below can also be introduced to enhance the model depending on the story-telling narrative we'll be adopting and how interpretable the final results are:

- Accuracy Score
- Precision
- Recall
- F1-Score
- Receiver Operating Characteristic

5. **Narrative formation and story-telling:**
   Finally, after all the predictions are done and the best possible models are chosen, we'll start the process of data visualization. Ideally several final 'decisive' features will be selected (or generated) and visualized into 2D/3D graphs for better understanding. Also feature-by-feature explainability effectiveness will also be visualized. Considering the complexity of this dataset, we'll try to reach global interpretability and local interpretability as well.
   Combined with the features' description provided, all the trends and global/local characteristics can be presented in an easily understandable way, through several representative cases that include distinctive features with wide enough variation.
   I.e:

   > An individual with *X* maximum delinquency ever, *Y* percent trades… will, according to our model, probably have a Good/Bad risk flag. Out of these information, X/Y/… contributed positively/negatively (by how much) to the final risk flag decision.

Conclusion:
We believe through the above process, we can gain a thorough understanding of the HELOC dataset, with high explainability and can produce a model that has an ideally high enough prediction score. Through illustrative graphs and well designed story-telling process these results can be clearly expressed and beautifully presented on the final day of project presentation.