**Student Bi-Weekly Performance Review**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Adm. No. | Name | No. of hours present | Progress1 | Remarks |
| 1. 2100775 | Soh Hong Yu | 8 | A | * Group Leader |
| 2. 2102719 | Rohan Ravishankar | 8 | A | * Nil |
| 3. 2136123 | Wong Tze Huai | 8 | A | * Nil |

**Bi-Weekly Scrum Report**

|  |  |
| --- | --- |
| Week No: 14-15 Date:16/01/2023 -27/01/2023 | |
| Member Name 1: | **Soh Hong Yu** |
| Last week’s Progress | * Decide on certain GUI design techniques  1. Buttons 2. Icons 3. Menus  * Research how to use DASK for model development and machine learning |
| This week deliverables | * Assists Tze Huai in Pipeline Development  1. Data preprocessing: Cleaning and preparing the data for analysis, including tasks such as handling missing values, removing outliers, and encoding categorical variables. 2. Feature engineering: Extracting and selecting the most relevant features from the data to use as input for the model. This can involve tasks such as feature scaling, normalization, and dimensionality reduction.  * Phase 2 report  1. Background research on how to indepthly expalin GUI. |
| Obstacles | * Decided to do pipeline development notebook quite late, thus not enough lead time |
| Member Name 2: | **Wong Tze Huai** |
| Last week’s Progress | * Machine Learning Model development  1. Problem definition: Clearly defining the problem 2. Model selection and training: Selecting a suitable machine learning algorithm and training the model on the prepared data. 3. Model evaluation and optimization: Evaluating the performance of the model on the test data and fine-tuning the model to improve its performance. |
| This week deliverables | * Model Selection * Model Optimization  1. Hyperparameter Tuning 2. Reularization 3. Ensemble Methods 4. Early Stopping  * Track and monitor machine learning cycles * Pipeline Development  1. Model selection: Choosing the appropriate machine learning algorithm for the task and tuning its hyperparameters. 2. Model evaluation: Evaluating the performance of the model on a test dataset, and using metrics such as accuracy, precision, recall, and F1 score to measure its performance. 3. Deployment: After finalizing the model, it can be deployed in production, where it can be used to make predictions on new data.  * Phase 2 Report  1. Do background research for in-depth report writing 2. Help Rohan with Introduction and layout of the report |
| Obstacles | * DASK takes very long to run and uses a lot of memory * Not used to DASK syntex |
| Member Name 3: | **Rohan Ravishankar** |
| Last week’s Progress | * Completed the comparison of different outlier detection using dask. **Found out that One-Class SVM is the best outlier detection. However, we will be using Local Outlier Factor as:**   1. SVM algorithm is not suitable for large data sets.  2. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.  3. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.  4. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.  5. SVM Fit time takes longer than LOF  Results:     1. Isolation Forest 2. Loacl Outlier Factor 3. One-Class SVM 4. Minimum Covariance Determinant 5. Robust Covariance  * Completed the comparion of different methods to balance classes. **Found out that Random Under Sampling is the best way to balance the classes.**   Results:     1. Random Under Sampling 2. Random Over Sampling 3. SMOTE 4. Tomek Links 5. ADASYN  * Found out that overall we should be using GradientBoostingClassifier for modelling as it has an overall better results for all outlier detection and balancing classes. |
| This week deliverables | * Scrum Report week 14-15 * Helped with the development of the pipeline.  1. Baseline Pipeline development  * Phase 2 Report Overview and Introduction * Help with Model Development Research * Re-do Data Aggregation to include Q1 & Q3 of the data |
| Obstacles | * Sprained ankle, could not come to school for 2 days * Some models took long to run * Had to rerun the models with new data so had took longer time than than expected. |