Page 1 of 20

Data Assurance

Provide text and/or evidence that support the following was followed as per the Model Card:

* appropriate data quality checks were performed
* feature selection process
* usage of data proxy, if applicable, was appropriate
* training/evaluation split methodology

**Summary**

During validation, no data anomalies were detected. The model pipeline includes data anomaly handling methods. Despite utilizing only 12 features, the validation process revealed that the model might perform equally well with even fewer features. However, this did not significantly impact the model's performance, and overfitting was avoided. The model was trained on a hand-labeled dataset, and as a result, the data was not partitioned into separate training and test sets.

**Verification:**

The total of 37,475 data points is used in this model. Due to the limited number of data the data was not split to training and test set. The data was hand-picked, and the labels were created manually. The distribution of the input data in time is shown in Figure 1. The data ranges from ?, to ?.

Figure 1- Distribution of the Input Data

The Data was controlled for duplicate and missing values as shown in Table 1. The distribution of the data is shown in Figure 2 across 4 clusters. Assurance on the hand labeled data.

Page 2 of 20

Table 1-Train Data Description

|  |  |
| --- | --- |
| Item | Train |
| Features | 12 |
| Samples | 37475 |
| Dimension | (37475,12) |
| Duplicated | 18789 |
| Missing Values | 0 |
| Negative Samples | 18691 |
| Positive Samples | 18784 |
| Balance (Negative/Positive) | 0.995049 |

Figure 2 - Test and Train Sample Distribution Across Clusters

As could be seen in the figure above, most of the samples could be clustered into a single cluster. That could be an indication of a potential skew in the sample. During the validation process we have also investigated the distribution of negative and positive cases in the data as shown in Figure 3. Another important aspect is the distribution of positive cases. We clustered the data and investigated the distribution of positive cases in each cluster across (Figure 4). As could be seen, most of the positive cases are in one cluster, which could be a result of the data distribution.

Page 3 of 20

Figure 3-Test and Train Positive and Negative Balance

Figure 4-Test and Train Positive and Negative Balance Across Clusters

To control the number of features, the validation process starts with performing a PCA. The PCA analysis suggest (Figure 5) that 2 orthogonal features may be able to explain a reasonable amount of variation in the data.

Page 4 of 20

Figure 5-PCA Analysis Results

Table 2-Sample of Pearson Score Results

|  |  |  |
| --- | --- | --- |
| feature1 | feature2 | Pearson correlation coefficient |
| sawtooth\_mt\_chng\_ratio4 | sawtooth\_mt\_chng\_ratio5 | 0.974417 |
| sawtooth\_mt\_chng\_ratio3 | sawtooth\_mt\_chng\_ratio4 | 0.933277 |
| sawtooth\_mt\_chng\_ratio3 | sawtooth\_mt\_chng\_ratio5 | 0.904777 |
| sawtooth\_mt\_chng\_ratio2 | sawtooth\_mt\_chng\_ratio3 | 0.803598 |

Table 3-Sample of VIF Score Results

|  |  |
| --- | --- |
| features | vif\_factor |
| sawtooth\_mt\_chng\_ratio3 | 0.974417 |
| sawtooth\_mt\_chng\_ratio4 | 0.933277 |
| sawtooth\_mt\_chng\_ratio5 | 0.904777 |

Table 2 presents the results of the Pearson correlation analysis, while Table 3 displays the same analysis with additional findings. In Table 4, the condition number of the features is listed, with values exceeding 30 indicating a relatively high condition number. Lastly, Table 5 exhibits the homogeneity score, highlighting 7 feature pairs with a homogeneity correlation coefficient exceeding 0.9.

Page 5 of 20

Table 4-Condition Number of the Data

|  |  |  |
| --- | --- | --- |
|  | eigenvalue | condition\_number |
| quote\_slope\_pos | 11573.85 | 202.576997 |
| sawtooth\_mt\_chng\_ratio1 | 144.73 | 1811.547693 |
| sawtooth\_mt\_chng\_ratio2 | 13.93 | 5839.201943 |
| sawtooth\_mt\_chng\_ratio3 | 4.13 | 10723.931410 |
| sawtooth\_mt\_chng\_ratio4 | 0.40 | 34458.713381 |
| sawtooth\_mt\_chng\_ratio5 | 0.30 | 39789.494893 |
| blkcontra\_blk\_rate | 0.08 | 77052.02537 |
| prop\_pr\_better\_qty\_ratio | 0.10 | 68917.426762 |

Table 5-Sample of Homogeneity Score Results

|  |  |  |
| --- | --- | --- |
| feature1 | feature2 | homogeneity correlation coefficient |
| odr\_wdth\_open\_qty\_ratio | sprd\_wdth\_open\_qty\_ratio | 0.998382 |
| sprd\_wdth\_open\_qty\_ratio | chain\_length | 0.997333 |
| odr\_wdth\_open\_qty\_ratio | chain\_length | 0.997283 |
| sawtooth\_mt\_chng\_ratio4 | sawtooth\_mt\_chng\_ratio5 | 0.947396 |
| sawtooth\_mt\_chng\_ratio3 | sawtooth\_mt\_chng\_ratio4 | 0.938736 |
| sawtooth\_mt\_chng\_ratio2 | sawtooth\_mt\_chng\_ratio3 | 0.925871 |
| sawtooth\_mt\_chng\_ratio3 | sawtooth\_mt\_chng\_ratio5 | 0.913252 |

The model is not using any scaling and is using the actual value. As the model is a random forest this is not that big of an issue as it is in deep learning models.

Was the data approach from the Model Card Followed?

Page 6 of 20

Model Selection & Training

Provide text and/or evidence that the Model Section Process was as correct including the following:

* hyperparameter tuning
* Training of the model
* feature selection process
* appropriateness of the model chosen

**Summary**

The model's performance is deemed satisfactory. However, it was observed during the validation process that the model lacks reproducibility.

**Verification:**

The intraday participant model adopts a Random Forest structure. To assess the model's reproducibility, the training pipeline was executed three times, and the resulting outputs were compared. The comparison of the confusion matrices is depicted in Figure 6, indicating that the model outputs are not identical. Moreover, Figure 7 illustrates the confusion matrix of the model. Additionally, Table 6 presents other goodness-of-fit measures. The model employs Greed Search to optimize the hyperparameters.

Figure 6-Comparison of (a) the First and Second Model (b) the First and the Third Model

Page 7 of 20

Figure 7- Model Confusion Matrix

The model, in general, successfully achieves its performance objectives as specified in the model card. For a comprehensive understanding and a complete list of model performance measures, refer to the detailed analysis. Figure 8 illustrates the distribution of false negatives within the dataset, while Figure 9 displays the model's goodness of fit across different clusters.

Page 8 of 20

Table 6-Goodness of Fit Measures for Test and Train Set

|  |  |
| --- | --- |
| index | score |
| accuracy | 0.994255 |
| recall score | 1.000000 |
| precision score | 0.988322 |
| f1\_score | 0.994127 |
| f2 score | 0.997642 |
| cohen\_kappa\_score | 0.988504 |
| roc\_auccuracy\_score | 0.994408 |
| false positive\_rate | 0.011183 |
| false negative\_rate | 0.000000 |
| true\_negative\_rate | 0.988817 |
| negative\_predictive\_value | 1.000000 |
| false\_discovery\_rate | 0.011678 |
| matthews\_corn | 0.988569 |
| avg\_precision | 0.988322 |
| log\_loss | 0.198443 |
| brier\_score\_loss | 0.005745 |
| binary\_ks\_curve | 0.988817 |
| balanced\_accuracy\_score | 0.994408 |
| top\_k\_accuracy\_score | 1.000000 |
| jaccard\_score | 0.988322 |

Page 9 of 20

Figure 8-Distribution of False Negatives in Test and Train Set in the Clusters

In order to gain deeper insights into the pattern of false negatives, we employed a decision tree to identify any potential rules that could help capture and mitigate them. The outcomes of this analysis are provided in [https://domino.finra.org/workspacesession/k31].

Since the model was trained on a single dataset, cross-validation was conducted to assess its robustness. The accuracies obtained from the cross-validation process are illustrated in Figure 10. The average accuracy of the model was found to be 99.5%, with a standard deviation of ?%.

Page 10 of 20

Figure 9-Model Goodness of Fit Measure (Precision and Recall) Across Clusters

Page 11 of 20

Figure 10-Ten-Fold Cross-Validation Results

The study of model robustness against adversarial attacks was also conducted. Each sample and feature were iteratively increased and decreased, and the resulting changes in the sample's label were recorded. Samples that exhibited label changes were identified as edge case samples. Overall, only a small number of samples were flagged as edge cases, except for blkcontra\_blk\_rate. The table below displays the features with the most significant contribution to the edge cases, which can be considered as sensitive features in this context. The findings are presented in Table 7.

The model was fed into sklearn to assess the feature importance. The results, shown in Table 8, demonstrate the significance of various model features. For in-depth analysis, PDP (Figure 11), ICE (Figure 12), ALE (Figure 13), SHAP (Figure 14), LIME (Figure 15), and Surrogate Model were utilized to evaluate feature importance. The comprehensive analysis is accessible on bitbucket [https://domino]. The figures below illustrate that sawtooth\_mt\_chng\_ratio2 is the most critical feature, while blkcontra\_blk\_rate is the least important feature in the model.

Page 12 of 20

Table 7-Features with Highest Sensitivity to Adversarial Attack

|  |  |  |
| --- | --- | --- |
| Type | Feature | Number of cases |
| UP | blkcontra\_blk\_rate | 2105 |
| sprd\_wdth\_open\_qty\_ratio | 3 |
| prop\_pr\_better\_qty\_ratio | 1 |
| sawtooth\_mt\_chng\_ratio1 | 1 |
| Down | blkcontra\_blk\_rate | 112 |
| prop\_pr\_better\_qty\_ratio | 49 |
| sawtooth\_mt\_chng\_ratio1 | 2 |
| chain\_length | 1 |
| sprd\_wdth\_open\_qty\_ratio | 1 |

Table 8- Most Important Features in the Model (Reverse Order)

|  |  |
| --- | --- |
| Feature | Importance |
| sawtooth\_mt\_chng\_ratio2 | 0.000121 |
| sawtooth\_mt\_chng\_ratio3 | 0.000181 |
| chain\_length | 0.000242 |
| sawtooth\_mt\_chng\_ratio4 | -0.000302 |
| opposite\_side\_price\_impact | 0.000484 |
| sawtooth\_mt\_chng\_ratio5 | 0.001572 |
| quote\_slope\_pos | 0.010463 |
| sawtooth\_mt\_chng\_ratio1 | 0.028092 |
| odr\_wdth\_open\_qty\_ratio | 0.035470 |
| sprd\_wdth\_open\_qty\_ratio | 0.133414 |
| prop\_pr\_better\_qty\_ratio | 0.181705 |
| blkcontra\_blk\_rate | 0.220774 |

Page 13 of 20

Figure 11-PDP Chart for the Most (a) and Least (b) important Feature of the Model

Figure 12-ICE Chart for the Most (a) and Least (b) important Feature of the Model

Figure 13-ALE Chart for the Most (a) and Least (b) important Feature of the Model

Page 14 of 20

Figure 14-SHAP Analysis of Model Features

Page 15 of 20

Figure 15-LIME Analysis of Model Features on Three Random Samples

Page 16 of 20

The current model was benchmarked against XgBoost as an alternative approach. The XgBoost model was trained using 80% of the data and evaluated on the remaining 20%. The performance of the model, without hyper-parameter tuning, on the test data is presented in Figure 16.

Figure 16 - Comparison of the Official Model with XEB

The Random Forest model was also trained with different sample sizes and compared with the rest of the data, as test set. The changes in the precisions show that the performance of the model would not deteriorate with 80% of the existing data (Figure 17).

Page 17 of 20

Figure 17-Effect of Sample Size on Model Performance

In addition, we conducted a comparison between the proposed model and the previous version. As depicted in Figure-18, the new model successfully eliminated some of the positive cases that needed to be removed. However, a noteworthy concern is that some negative cases are now being classified as positive.

Figure 18-Comparison of the Existing Model and the Previous Model

Was the Model Selection approach from the Model Card Followed?

Page 18 of 20

Model

Provide text and/or evidence that the Model Monitoring Plan is satisfactory. Include how it ensure that the model continues to perform as expected in accordance with its design objectives and business purpose.

**Summary**

No modifications were made to the monitoring plan, and the new version of the model will adhere to the previously approved monitoring plan without any alterations.

Is the Proposed Model Monitoring Plan satisfactory? yes

Page 19 of 20

Assurance Goals

Provide test and or evidence that the Assurance goals have been met including the following:

* evidence of the execution of the Assurance Plan
* built and tested as per plan
* Provide results of the model
* Quantitative Analysis (Perf over total data set and subsets of data)
* Model Monitoring Plan is sufficient

**Summary:**

Describing high-level what Assurance Goals were met and what Assurance Plan was followed.

**Verification:**

The assurance plan is located on confluence page:

[<https://wiki.finra.org/display/MarketRegulationSurveillance/Intraday+Assurance+Plan>]

Was the type and level of Assurance activities performed sufficient?

Page 20 of 20

Provide text and/or evidence that the model is meeting its performance and appropriateness including its design and business objectives.

**Summary:**

In summary the model meets its success criteria as defined in the model card and improves the previous version of the model in terms of identifying out of window market activities.

**Verification:**

The previous model was not able to capture the out of the window market activities. The new version of the model aims at targeting and identifying these activities and add them to the training and prediction model. This ensures that the window of interest is not just the typical market activity.

The success criteria defined for this retraining effort is a potential 50% reduction is positive cases from the previous model. It has been verified that compared to the previous version of the model, the retrained model is removing 60% of the positive cases.

Does the model meet the identified design objectives?

Does the model meet the identified business purposes?