Model Validation Report

[DOCUMENT SUBTITLE]

**Document Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Description of change | Author | date |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Model Info**

|  |  |
| --- | --- |
| Model | Intrusion |
| Model Version | 1.0 |
| Date |  |
| Primary Validator | Navid Kalantari |
| Model Registry link |  |

**Scope**

The following encompasses the scope of the MVR:

* An independent assessment of the appropriateness & performance of a model to meet its design objectives
* Observations of technical model risks and/or limitations, with associated mitigation steps taken
* Business risks associated with the model usage and the mitigations to those risks
* Model validator sign off that required and specified procedures for developing model were followed
* Summarize results of executing assurance plan
* Summarize other mode validation results

مواردی که هایلایت شده رو قبلا جیمز نوشته بود برای ولیدیشن خودش

**Model Performance/Appropriateness**

**Summary:**

FDID's that participate In Intrusion events are flagged for Individual days. There are currently three-steps:

1. (Deep-learning based: Supervised) A shape model that detects instances of suspicious looking shifts in option price and particularly low delta values.
2. (Deep-learning based: Unsupervised): An autoencoder model that takes positives from the shape model and scores them based on how unusual they appear to look.
3. (Rules-based): A participant model that looks at how high-score results from the autoencoder, and flags which FDID's have a proportion (determined by a pre-specified 'purity filter) of the high-score results that exceed a set of pre-defined value thresholds.

The end model gives a set of output that includes all high-score instances from flagged FDIDS, regardless of value thresholds.

Performance generally will be assessed by usefulness of positively labeled instances to business users. Based on the feedback from users during the prototyping process, we expect the current iteration of the model will produce potential instances of intrusion that should be of interest to users about 90% of the time. The visual output of the model will allow users to quickly determine whether or not a flagged instance is worthy of further investigation.

**Verification:**

Check dashboard: The pattern produces plots and data in SURF which can be reviewed by business users.

لطفا دشبورد رو به عنوان خروجی چک کنین.

Model is Bayesian, below shows rates for each set at thresholds of 0.6 and 0.8-meaning a prediction is considered a positive overall if it is labeled positive for 60 out of 100 or 80 out of 100 Iterations respectively.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data Type | Bayesian coutoff | True Negatives | False  Negatives | True Positives | False  Positives | Precision | Recall | Accuracy | F-Measure |
| Value | 0.8 | 215185 | 1035 | 545 | 31 | 0.946181 | 0.344937 | 0.995083 | 0.505566 |
| Test | 0.8 | 117930 | 1102 | 212 | 34 | 0.861789 | 0.161339 | 0.990476 | 0.271795 |
| Train | 0.8 | 850398 | 4648 | 1811 | 80 | 0.957694 | 0.280384 | 0.994483 | 0.433772 |
| Value | 0.6 | 214884 | 1336 | 566 | 10 | 0.982639 | 0.297581 | 0.993791 | 0.456420 |
| Test | 0.6 | 117628 | 1404 | 219 | 27 | 0.890244 | 0.134935 | 0.988003 | 0.234350 |
| Train | 0.6 | 849078 | 5968 | 1844 | 47 | 0.975145 | 0.236047 | 0.992961 | 0.380089 |

What processes were used to handle imbalance during training such as Under sampling or SMOTE? SMOTE was used to upsample positives for an equal balance of positives and negatives during the training process.

اینجا گفته شده که مدل بالانس شده ولی نتایج کانفیوژن ماتریس میگه که بالانس نشده اند

Autoencoder (AC):

|  |  |  |
| --- | --- | --- |
|  | AC Negative | AC Positive |
| Label Negative | 3100000 | 670000 |
| Label Positive | 720 | 5040 |

Balance check:

Does the model meet the identified design objectives? 🞏

Does the model meet the identified business purposes? 🞏

اگر اهداف حاصل شده لطفا این دو مورد رو تیک بزنین

**Data Assurance**

**Summary**

The data used to train the model is from POD and OCC, and a table of Intrusion Instances found in namespace object. Trade data is only Included from days where there is at least a single labeled Instance of intrusion behavior, and likewise, positive instances are only included on days where POD data could be successfully pulled. (There are 4 such days where POD days produced errors). Data was pulled from the year 2021 to train the model.

The features selected for the model have either been identified by MRMR (maximum relevance, minimum redundancy) out of thousands of candidates, or have been selected for principled reasons.

Feature selection was performed using MRMR, with the optimal number of features being assessed by analyzing how performance with three predictive measures Increased as set-size Increased.

Medians for both the 'delta' value of each rolling window, and the meidans of the four types of data used for feature selection (bid-ask spread, trade-price bid spread, option ask, and option bid), were all incorporated.

What is the 'delta' value?

The data has not been investigated for outliers-visual Investigation of scaled features suggests that extreme/skew values in many cases may be very helpful for prediction. All features are scaled within a range of 0 to 1, so this should limit any undue Influence. Future Iterations of this model may however, find that a more sophisticated treatment of outliers/skew more generally may improve results.

دلتا ولیو نمیدونم چی هست؟

Shape model training Involves splitting the data by symbol into training (75%), validation (15%), and test (10%) buckets. Symbols are selected randomly and placed into buckets, percentages reflect total number of rolling windows (not total number of symbols

|  |  |  |
| --- | --- | --- |
| **Set** | **Number Positives** | **Number Negatives** |
| **Training** | **2300** | **1100000** |
| **Validation** | **500** | **200000** |
| **Test** | **330** | **150000** |

The number of sets doesn't match with the confusion matrix.

Ground truth data of intrusion incidents is provided in the OPT\_OPTINTRU\_XCPTN and opt\_optintru\_dtl\_xcptn\_v1, no proxy is needed for positives, and for the purposes of trading all other option trades in the appropriate time frame are assumed negative.

**Verification:**

Provide any evidence that the data approach was followed

دیتا و کیفیت آن نیاز به چک و تأیید کردن شما دارد.

training/evaluation split methodology??

Was the data approach from the Model Card Followed? 🞏

اگر اهداف حاصل شده لطفا این مورد رو تیک بزنین

**Model Selection & Training**

**Summary**

Hyper-opt was suppressed during the training of the most recent model.

One hidden dense variational layer of 20 nodes was found to be sufficient to generate very high value AUC scores. 150 Epochs were used. Batch size was set to 2048

Success was found in the most recent model by varying the learning rate, running one iteration at 0.03, another at 0.003 followed by 0.0003.

No tuning process was used in the most recent iteration of model fitting, very high model validation scores were achieved with a straightforward set of hyperparameters.

How does the performance of the model on the test data compare to the benchmark performance on the same data? This has not been evaluated.

What is model performance at extreme ranges of input data? This has not been tested to date

What was the impact of small changes to input values? Does the model output vary within expected ranges? This has not been tested to date

موارد هایلایت به نظرم به این بخش ربط داشت

**Verification:**

Was the Model Selection approach from the Model Card Followed? 🞏

اگر اهداف حاصل شده لطفا این مورد رو تیک بزنین

**Model Monitoring Plan**

**Summary**

Describe how the model is monitored. Can the users verify whether the model is running in a self-service method? Is there a dashboard or report that users can access to review the latest monitoring? Does the monitoring allow users to review the model output by different slices or groups of the data?

Daily alerts/reports/dashboards will be available to users in Surf.

**Verification:**

Is the Proposed Model Monitoring Plan satisfactory? 🞏

اگر اهداف حاصل شده لطفا این مورد رو تیک بزنین

**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

این موارد رو نتونستم پیدا کنم!

**Summary:**

The deep-learning pattern is intended to use tsfresh features to detect instances of intrusion behavior that may not be detected by purely rules-based solutions. This will identify a broader set of cases of intrusion behavior, and also may help business users identify specific FDID's to target.

This model was selected to increase the number of instances detected, the combination of ts-fresh data and neural networks are particularly strong at detecting shapes in the movement of prices and spreads, and locating them at scale. The autoencoder was included to reduce the number of false positives.

**Verification:**

Click or tap here to enter text.

Was the type and level of Assurance activities performed sufficient? 🞏

اگر اهداف حاصل شده لطفا این مورد رو تیک بزنین

**Risks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk Type (Business/Technical/Other)** | **Risk Description** | **Severity** | **Mitigation** | **Potential Impact** |
| Business Risks | Thresholds and filtering in AGR portion of model could lead to under-detection of certain types of intrusion behavior. |  |  | Medium |
| Technical Risks | Issues with pulling POD data in a timely matter and at scale could make retraining time consuming. |  |  | Low |
| Technical Risks | MRMR does not select consistent features with different sets of input data, suggesting process is very sensitive to noise. Predictions still consistent. |  |  | Low |

After feature values are scaled from 0 to 1, missing values are given a score of 0.5

All NaN's are replaced with 0.5 after scaling.

No tuning process was used in the most recent iteration of model fitting, very high model validation scores were achieved with a straightforward set of hyperparameters

How does the performance of the model on the test data compare to the benchmark performance on the same data? This has not been evaluated.

What is model performance at extreme ranges of input data? This has not been tested to date

What was the impact of small changes to input values? Does the model output vary within expected ranges? This has not been tested to date

موارد هایلایت به نظرم میتونن ریسک باشن ولی وارد نکردم که شما تصمیم بگیرین براشون

**Sign Off**

Provide a formal sign off on the release of the model under test.

Click or tap here to enter text.