Ergebnisse Masterarbeit

Christoph Großauer – Reliability Process

# Input Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Min | Max | Example | Perturbation Note |
| DEP\_DELAY(MINS) | -188 | 181 | 107 | Only integers allowed |
| CRS\_ELAPSED\_TIME(MINS) | 45 | 530 | 89 | Only integers allowed |
| NR\_PREV\_ARR\_FLIGHTS(1HR) | 0 | 120 | 74 | Only integers allowed |
| ARR\_DAY\_SIN | -0.975\* | 0.975\* | -0.434\* | Be aware of resolution to next day |
| ARR\_DAY\_COS | -0.901\* | 1 | -0.900\* | Be aware of resolution to next day |
| ARR\_MIN\_OF\_DAY\_SIN | -1 | 1 | -0.985\* | Be aware of resolution to next day |
| ARR\_MIN\_OF\_DAY\_COS | -1 | 0.999\* | 0.169\* | Be aware of resolution to next day. |
| APPROACH\_SPEED(KMH) | 211.128 | 290.764 | 255.576 | Average gap of +-0.7% to next value overlap. Perturbate within this margin. |
| TAIL\_HEIGHT(M) | 6.325\* | 19.583\* | 9.001\* | Average gap of +-2% to next value overlap. Perturbate within this margin. |
| PARKING\_AREA(SQM) | 527.871\* | 4500.743\* | 1525.177\* | Average gap of +-3% to next value overlap. Perturbate within this margin. |
| WINGLETS(YN) | 0 | 1 | 1 | No values other than 0 and 1 |
| TEMP(C) | -9.389\* | 34.389\* | 6.722 | +-0.75% tolerance due to 0.1°F resolution. Calculated on median temperature impact. |
| DEWPOINT\_TEMP(C) | -15 | 24.389 | -3.278 | +-0.9% tolerance due to 0.1°F resolution. Calculated on median dew point temperature impact. |
| REL\_HUMIDITY(PERCENT) | 12.7 | 100 | 48.88 | +-2% tolerance based on worst case scenario on calculation with tolerances of temperature fields. (Since humidity is calculated based on them) |
| WIND\_DRCT(DEG) | 5 | 360 | 290 | 5-degree gaps within measurements. |
| WIND\_SPEED(KMH) | 0 | 62.042 | 26.854 | +-3,7% tolerance or 5%. Whichever is bigger. |
| 1HOUR\_PRECIPITATION(INCH) | 0 | 0.915 | 0 | +-0.02 inches tolerance |
| SEA\_LEVEL\_PRESSURE(MILLIBAR) | 987.1 | 1041.2 | 1018.8 | +-0.7 millibars tolerance |
| VISIBILITY(MILES) | 0 | 10 | 10 | +- 1 tolerance possible. |
| EVENT\_BR | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_DZ | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_FG | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_FU | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_GR | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_GS | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_HZ | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_IC | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_RA | 0 | 3 | 0 | No -1 allowed |
| EVENT\_SN | 0 | 2 | 0 | No 3 found in train set. |
| EVENT\_TS | 0 | 3 | 0 | No -1 allowed |
| RUNWAY\_ERROR(PERC) | 0 | 1 | 0.4 | Only 0.2 steps in data |

\*… Value is rounded

# Model Explanation and Input Data

The current best performing model, which is further used for the perturbation analysis is a Random Forest Classifier from the sklearn.ensemble module. Because it is a Random Forest Classifier we do not need to make use of over- and undersampling techniques, since the model has a parameter to account of class imbalance. (class\_weight: balanced)

The above input data is fed into the model as described for training purposes. The data was split prior to any operations regarding data alterations which also means, that the model might encounter values out of the described range on the evaluation phase. This is where perturbating out of variable ranges might yield interesting solutions.

The target distribution for the model’s train and test set is described in the following image. (Early < -15 min; -15 min <= On Time <=15; 15 < Late)

Ein Bild, das Screenshot, Diagramm, Grafiksoftware, Multimedia-Software enthält.

Automatisch generierte Beschreibung

The distribution stays intact and shows the imbalanced problem at hand.

Experiments with the input data has also shown, that scaling is not necessary for the performance of the Random Forest Classifier trained in this use case. Therefore, the input values are preserved throughout the whole pipeline.

To find a near-optimal parameter set for the model, Exhaustive Grid Search with 5-fold cross validation was conducted.

## Evaluation

Ein Bild, das Text, Screenshot, Zahl enthält.

Automatisch generierte Beschreibung

Above, the confusion matrix for test-set evaluation is shown. Especially label ‘2’ entries are recognized at a reasonable rate with 73% while also still capturing lots of the on-time flights. Prediction of early-arrivers is not optimal, but separating early and on-time flights was hard to achieve, since most flights arrive a little early within the dataset as depicted below.

Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung

We see that the delays are strongly skewed towards the -15 minutes threshold. This skew towards early arrivals makes correct classifying on this side of the dataset harder than on the +15 mins delay side.

Other notable evaluation metrics are shown in the tables below:

|  |  |
| --- | --- |
| **Global Measures** | |
| Accuracy | 73,43% |
| Makro Precision | 75,10% |
| Makro Recall | 71,57% |
| Makro F1 | 73,03% |

|  |  |  |  |
| --- | --- | --- | --- |
| **One vs. All Measures** | | | |
|  | Early | On Time | Late |
| Precision | 58,30% | 78,22% | 88,78% |
| Recall | 63,96% | 77,84% | 72,91% |

### Baseline values

#### Random Guessing

Two baselines have been chosen to evaluate the project. The first baseline randomly guesses the target class according to the training distribution of the target. 30% early, 56% on-time and 14% late.

Ein Bild, das Text, Screenshot, Zahl enthält.

Automatisch generierte Beschreibung

#### Delay Forwarding

This baseline simply uses the column DEP\_DELAY and 1:1 transforms the value onto the target and bins the result to match the classification scenario.

This baseline acts similar to a passenger on a plain who estimates his arriving delay by simply adding the departure delay to the planned arrival.

Ein Bild, das Text, Screenshot, Zahl enthält.

Automatisch generierte Beschreibung

#### Metrics of Baselines

|  |  |  |
| --- | --- | --- |
| **Global Baseline Measures** | | |
|  | Random Guess | Delay Forwarding |
| Accuracy | 43,02% | 64,17% |
| Makro Precision | 33,45% | 75,08% |
| Makro Recall | 33,45% | 57,56% |
| Makro F1 | 33,45% | 51,11% |

Both baselines were beaten by the model in all regards. Although makro precision of the Delay Forwarding baseline comes close to beating the model, the advances in the other metrics are sufficient for the project.

## Perturbation

As a first step we try to investigate the varying perturbation levels. Since the lvl 2 perturbation options did not alarm any entry, we only have 2 levels to check. The following table shows the amount of perturbation options per level, the number of alerts coming from each level and the according percentages.

Ein Bild, das Text, Schrift, Software, Multimedia-Software enthält.

Automatisch generierte Beschreibung

The level 2 options in this scenario are checking for events to be out of bounds. For example, having a value of 3 even though only 2 is allowed as a maximum.

The following table shows the number of alarmed options per column name. We see that DEP\_DELAY is a very sensitive column, which is to be expected, since it is such a vital part of the dataset. 1HOUR\_PRECIPITATION is possibly problematic, since the sensor measurement is not precise.

Ein Bild, das Text, Screenshot, Software, Multimedia-Software enthält.

Automatisch generierte Beschreibung.

### Graphical Analysis

First, I went through graphical analysis of the perturbation results in which I try to compare distributions of the data and compare it with the distribution of all data entries that lead to a perturbation alarm according to the perturbation options. This approach might find interesting relationships of the model regarding corner cases and classification margins by observing peaks or underrepresented value ranges in the distribution.

Interesting findings for the graphical analysis were:

Ein Bild, das Screenshot, Multimedia-Software, Digitales Compositing, 3D-Modellierung enthält.

Automatisch generierte Beschreibung

Although most values are found around the 10 km/h mark, the perturbated values are concentrated on the upper end of the distribution. Potential indication, that the model is more sensible on higher wind speeds than it is on lower speeds. Additionally, wind speed has a high percentual tolerance (3.7%) which is most likely the reason for the skew to the higher end of the distribution. Operators of the model should therefore be aware that higher wind speeds are not as accurate and may lead to wrong classifications.

Ein Bild, das Screenshot, Spielesoftware, Digitales Compositing, PC-Spiel enthält.

Automatisch generierte Beschreibung

Wind direction is an example of overlapping distribution. Although wind direction of 0 degrees and around the 80-degree mark seems to be more sensitive to change.

Ein Bild, das Screenshot, Digitales Compositing, Spielesoftware, 3D-Modellierung enthält.

Automatisch generierte Beschreibung

One of the most obvious examples for anomalies in the distribution is the number of previous arriving flights (Traffic measure). Entries with flights below the 40 mark are not changing perturbation results at all, which makes this column very robust in the lower-end.

Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung

Very similar to the previous arriving flights, the dewpoint temperature seems to be robust on the lower end.

Ein Bild, das Screenshot, 3D-Modellierung, Digitales Compositing, Multimedia-Software enthält.

Automatisch generierte Beschreibung

CRS\_ELAPSED\_TIME has a high number of overall perturbation alarms with a peak of over 400 hits around the 125 mins mark, which makes it a very sensitive column.

### Evaluation Metrics

Additional to the graphical analysis, we observed changes on traditional metrics like accuracy, precision and recall when only looking on subsets of the data. The idea being, that the perturbation approach might allow people to identify possible bad entries with higher possibility of misclassification through dividing the dataset in a portion of unchanged entries and alarmed entries.

In the next screenshot we investigate the accuracies for each level, to see if perturbation isolation might help improving accuracies of the subset.

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

Looking at all perturbation options as shown above, seems to allow us to pick out worse performing areas of the dataset. Accuracies are generally higher in the non-alarming groups.

Additionally, we apply the same logic as above onto the test data such that we can check if a given entry in the test data can be isolated by using the perturbation approach. For this we isolate alarmed values and check the impact the isolation has on traditional evaluation metrics.

Ein Bild, das Screenshot, Grafiksoftware, Multimedia-Software, 3D-Modellierung enthält.

Automatisch generierte Beschreibung

The red line indicates the possible margin of global accuracy or label-based precisions. The bar indicates the “real” values when the whole test set is taken into account. The lower end of the red bar is the score when only data entries which are covered by some perturbation alarms are looked at. The higher end represents the score when all alarmed entries would get discarded and only the “robust” entries remain in the set. The exact values for this are also represented in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision 0 | Precision 1 | Precision 2 |
| Real Test Results | 73,43% | 58,30% | 78,22% | 88,78% |
| Only Alarmed | 60,43% | 52,64% | 68,09% | 55,02% |
| No Alarmed | 84,16% | 71,97% | 84,40% | 94,85% |

In theory, we could further improve these values by assuming that someone with domain knowledge could fix the alarmed values. This would reduce the number of entries that are to be manually checked drastically and could in theory boost all measures immensely.

#### Possible measures when all alarmed entries are fixed:

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision 0 | Precision 1 | Precision 2 |
| 91,32% | 90,55% | 90,81% | 95,88% |

Of course, this is highly unrealistic but gives some perspective on the true upper limit of the perturbation in terms of metric benefit.

We could also assume that a manual operator misclassifies all alarmed entries which would yield a lower limit of this approach

#### Possible measures when all alarmed entries are misclassified:

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision 0 | Precision 1 | Precision 2 |
| 73,09% | 57,85% | 77,84% | 88,69% |

### Local Accuracy

As a last experiment I tried to investigate local accuracy and checking whether alarmed entries also do have lower metrics than non-alarmed entries. Due to the dimensionality of the data, I used Manhattan-Distance and looked at the 4 closest neighbours (Chosen by empirical experiments) and calculating the mean of all alarmed and non-alarmed entries.

Unlike the other experiments, this could not be replicated consistently over various models and parameter combinations and was therefore deemed as unsuccessful.

## Conclusion

The reliability checking approach can impact machine learning projects in various ways. First, it promotes good documentation and business understanding habits. Furthermore, data inconsistencies, preprocessing steps and sensory tolerances are captured in a structured way.

Additionally, it may help to find problematic predictions and locate weak areas within the data structure which could help to improve the runtime performance by human intervention.