openIMIS

AI Claim Categorization as a Global Good

Simona Dobre, Dragos Dobre, Siddharth Srivastava

Gumzo ya Mwezi Montly Meeting

04 April 2022

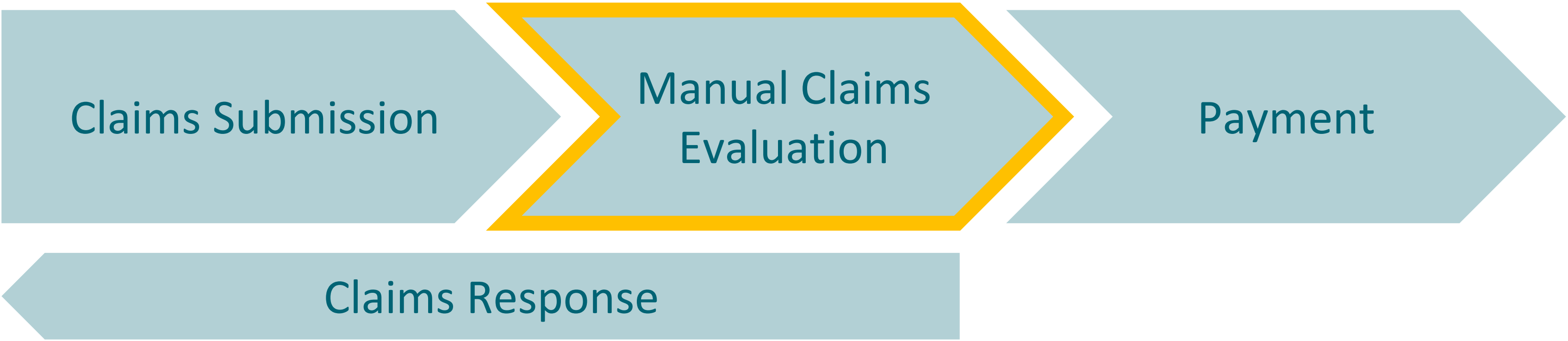
Content

- Claim Adjudication
- AI-based Claim Adjudication
- Framing questions

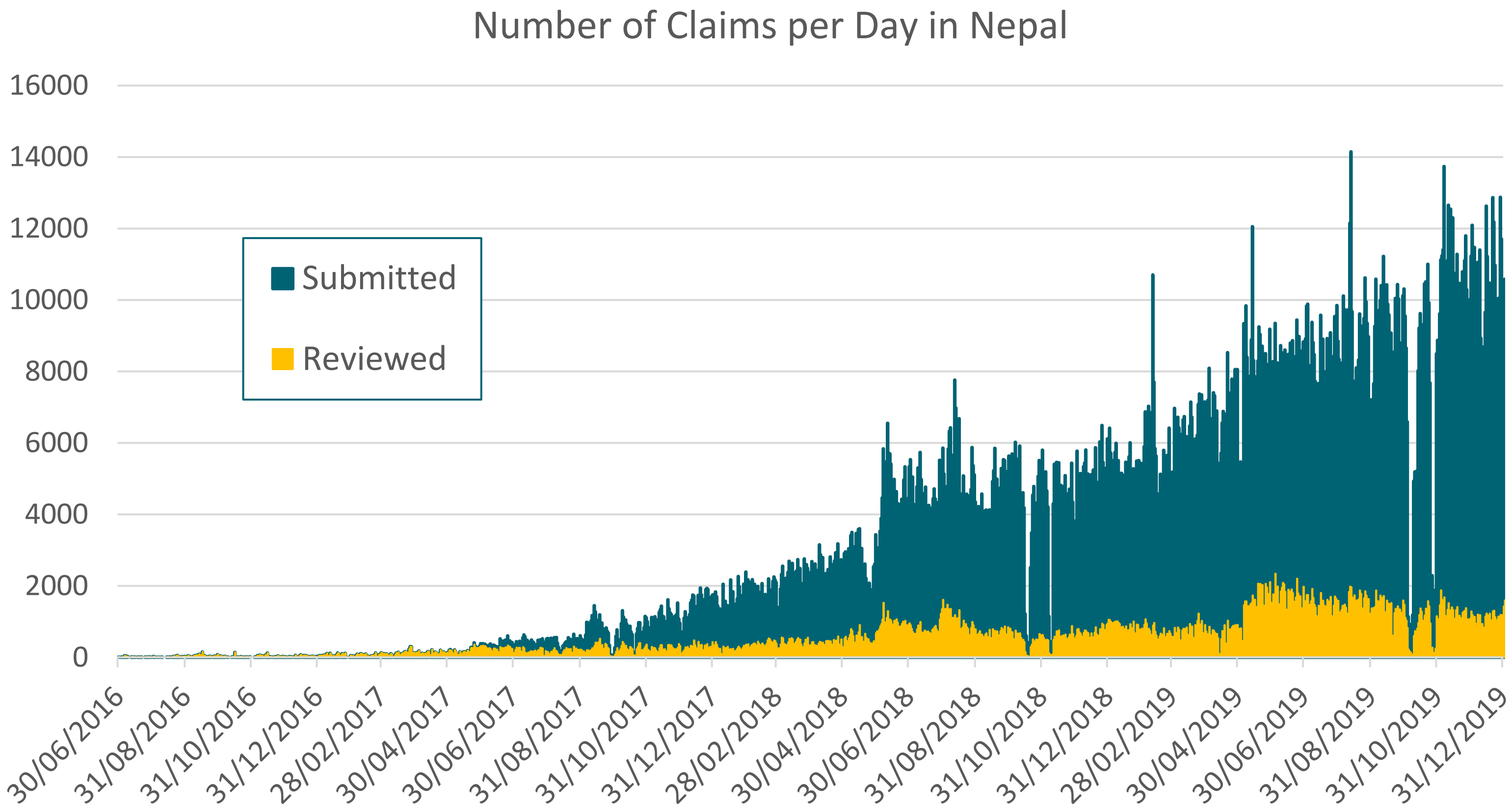
openIMIS

Claims Adjudication

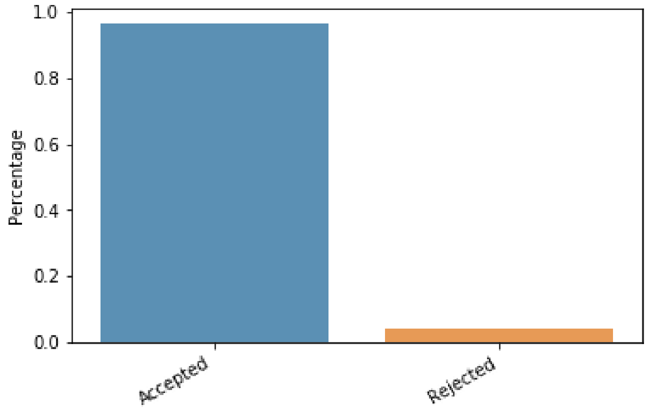
Manual Claims Adjudication



Problem: Number of Claims



- 5 953 640 Claims:
  - 12 371 992 Medical Items
  - 16 655 364 Medical Services
- 3 790 789 Insurees
- 780 Health Facilities



Bottelneck: Human Adjudicators

- Estimation (openIMIS Nepal):
  - 1 officer = 100 claims per day max
- Currently employed:
  - 16 officers = 1,600 claims per day
- Needed:
  - 300 officer = 30,000 claims per day

openIMIS

Adding AI

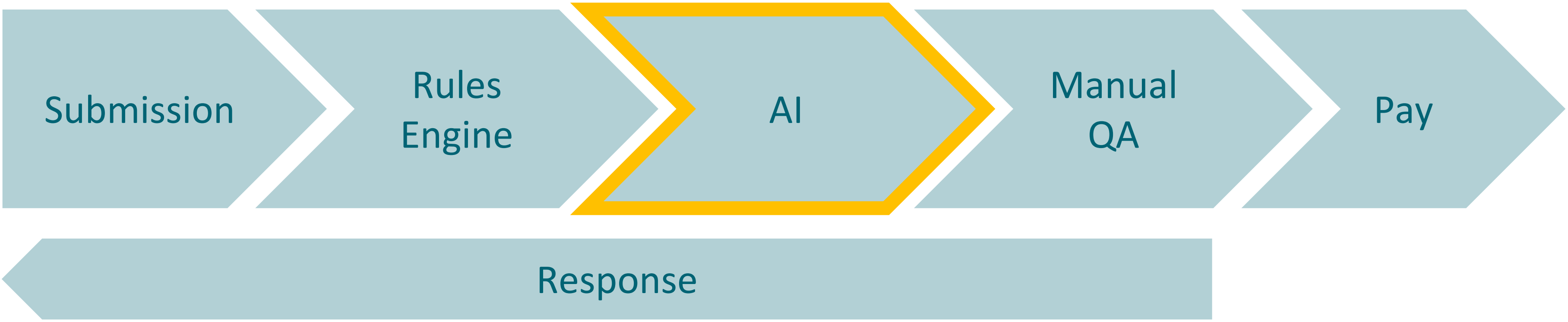
Rule Based Automation

Aim:
- Reduce workload

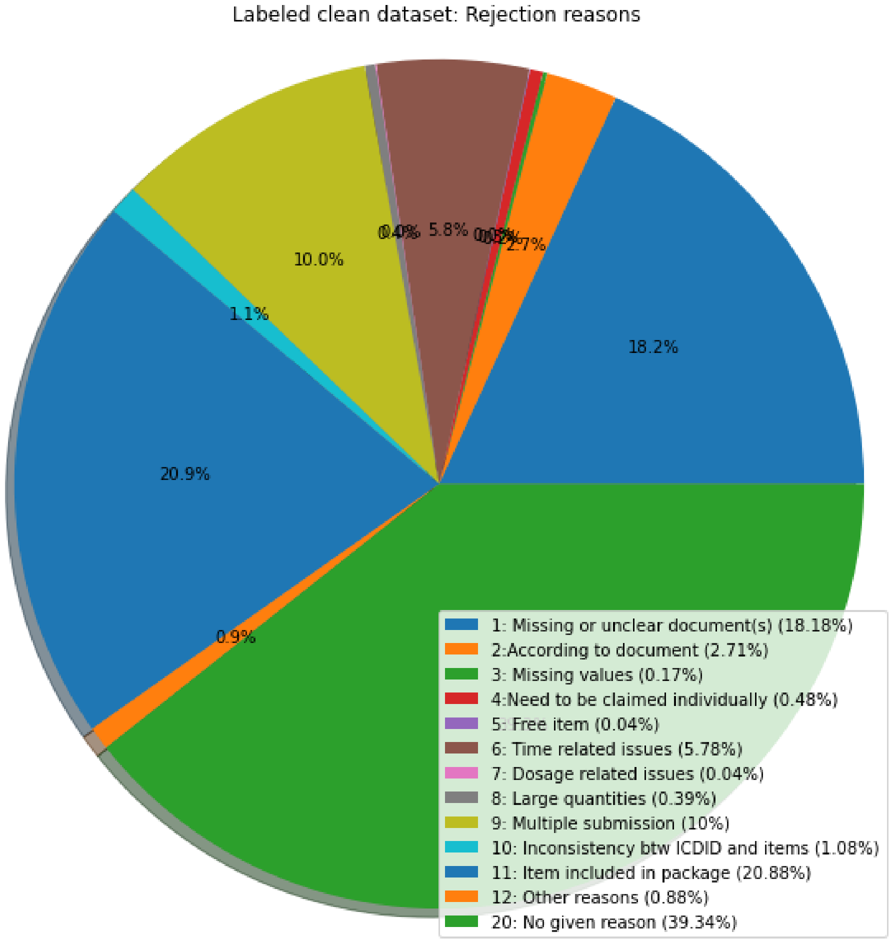
Method:
- Automatically reject formally incorrect claims
- No manual verification of rejected claims

Formally this is already Artificial Intelligence (but not Machine Learning and not the hype thing)

AI Supported Automation



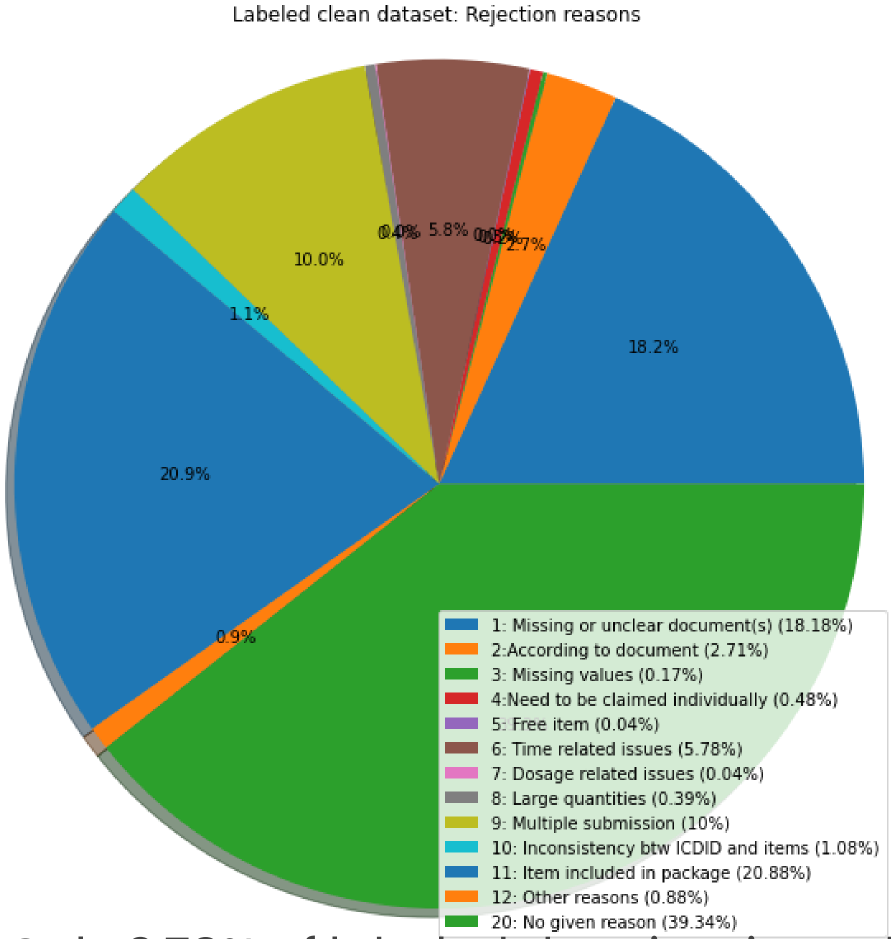
Challenges on data analysis



- Only 3.78% of labeled data is rejected, while only 2.29% has an associated rejection reason
    - Rejection justifications are free non standardized text fields ⇨ need to **process this information** in order to extract rejection reasons and standardize the Justification/Adjustment field
    - ⇨ dealing with **highly imbalanced dataset**
- Most of the features are categorical

⇨ only **specific AI models** are capable to consider this

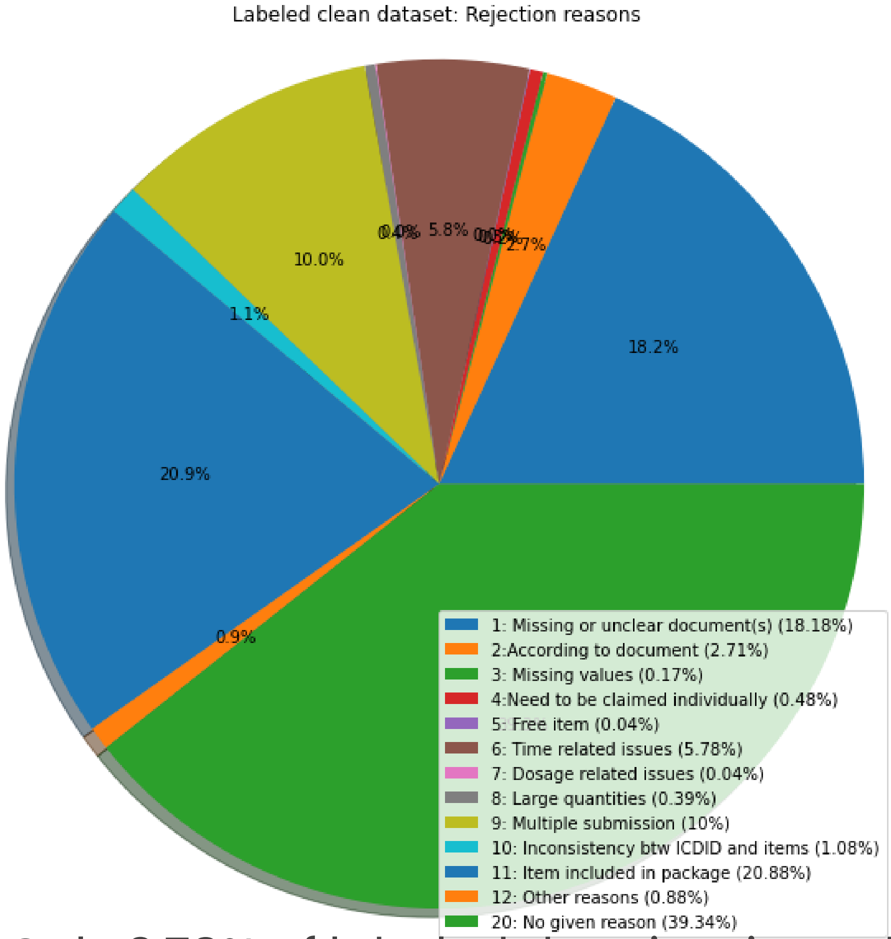
Challenges on data analysis



Only 3.78% of labeled data is rejected, while only 2.29% has an associated rejection reason

- Most of the features are categorical
  - Numerical: QuantityProvided, PriceAsked, ItemPrice
  - Categorical:
    - Date: DateFrom, DateTo, DateClaimed, DOB
    - Related to categories: ItemFrequency, ItemPatCat, ItemLevel, VisitType, HFLevel, HFCareType, Gender, ItemServiceType, PovertyStatus
    - ID related: ItemID, ClaimID, ClaimAdminID, HFID, LocationID, HFLocationID, InsureeID, FamilyID, ICDID, ICDID1

Challenges on data analysis

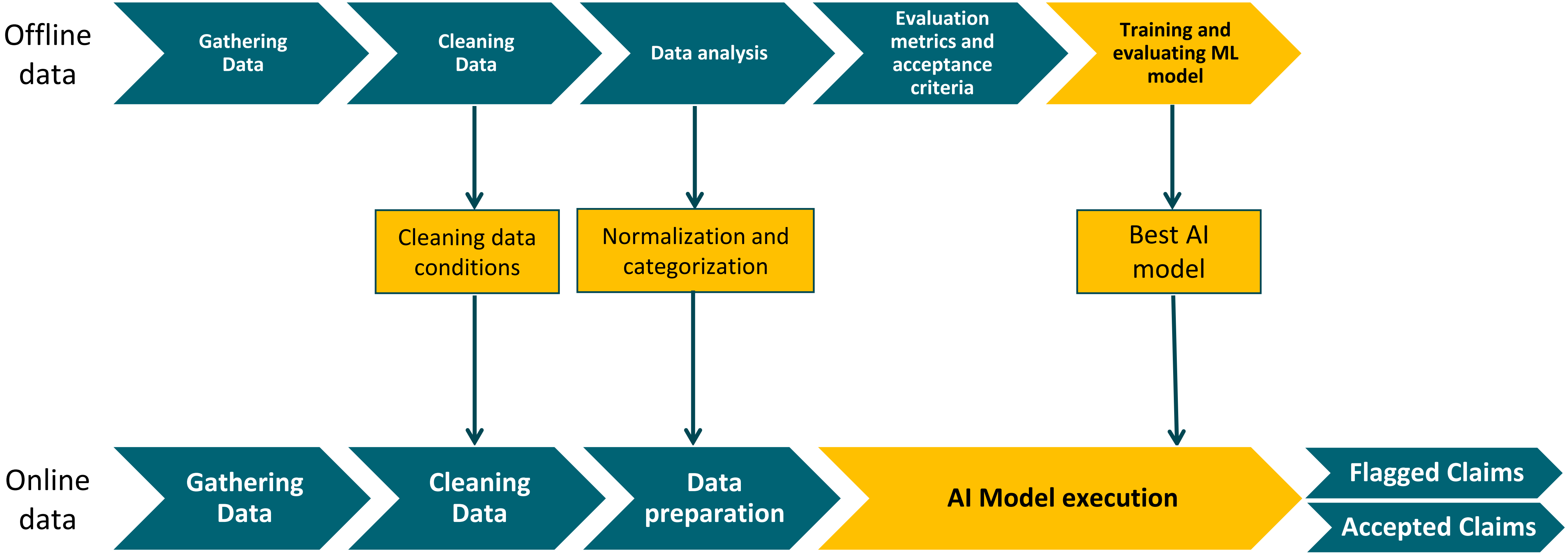


Only 3.78% of labeled data is rejected, while only 2.29% has an associated rejection reason

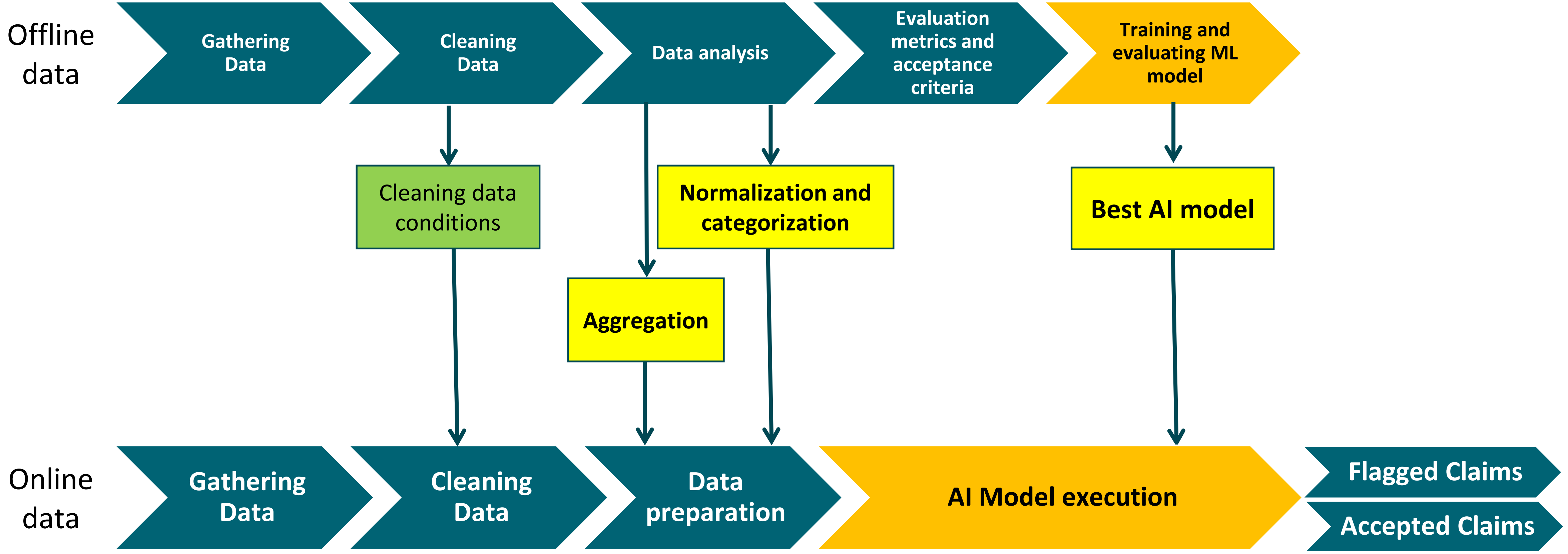
- Most of the features are categorical
  - Numerical: QuantityProvided, PriceAsked, ItemPrice Duration, DurationClaimed, Age
  - Categorical:
    - Date: ~~DateFrom, DateTo, DateClaimed, DOB~~
    - Related to categories: ItemFrequency, ItemPatCat, ItemLevel, VisitType, HFLevel, HFCareType, Gender, ~~ItemServiceType~~, Poverty
    - ID related: ItemID, ClaimID, ClaimAdminID, HFID, LocationID, HFLocationID, InsureeID, FamilyID, ICDID, ICDID1

⇨ replace ID related fields with aggregated fields

From research to production



From research to production



Excluding conditions

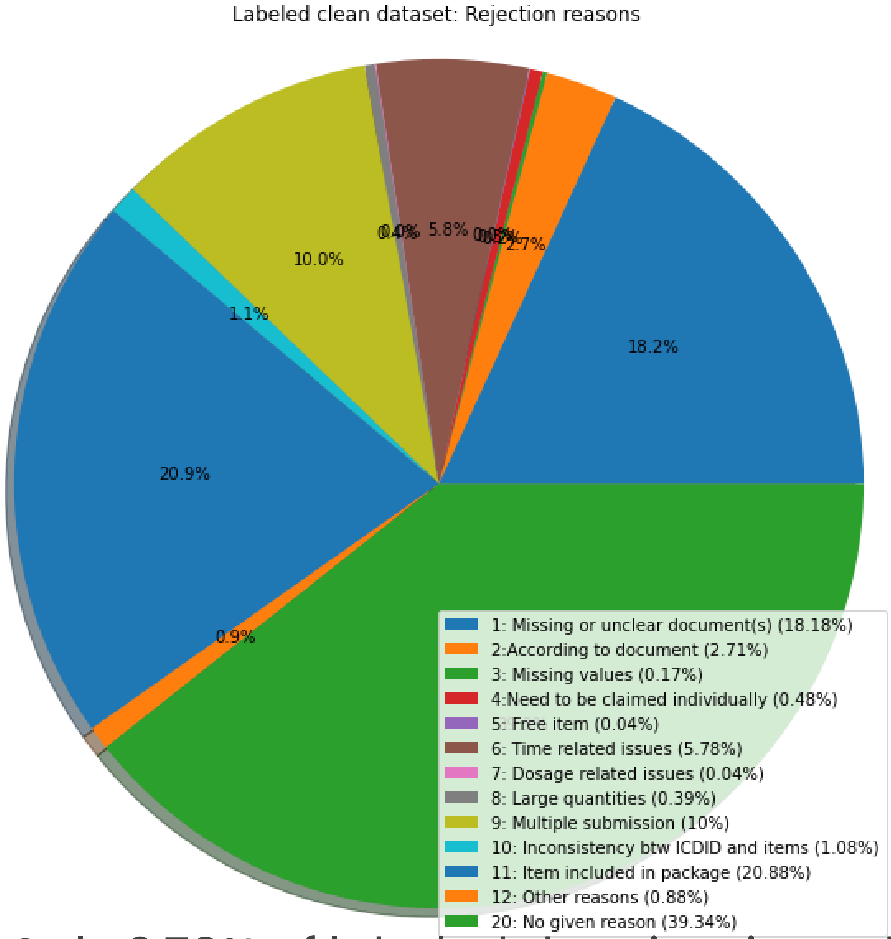
<table>
<tr><th>Condition</th><th>Description</th></tr>
<tr><td>1. df[‘ClaimStatus’] == CS_Entered</td><td>The items that are submitted, but not yet in checked by the Rule Engile are excluded</td></tr>
<tr><td>2. df['RejectionReason']>RR_Accepted</td><td>Items rejected by the Rule Engine are excluded</td></tr>
<tr><td>3.(df['RejectionReason']==RR_RbyMO)&(df['PriceValuated']>0)</td><td>Incoherence between status and valuated price</td></tr>
<tr><td>4. ((df['ClaimItemStatus']==CIS_Rejected)&\
(df['RejectionReason']==RR_Accepted))|\
((df['ClaimStatus']==CS_Rejected)&(df['RejectionReason']==RR_Accepted))|((df['ClaimItemStatus']==CIS_Accepted)&(df['RejectionReason']==RR_RbyMO))</td><td>Incoherence in the status fields are excluded</td></tr>
<tr><td>5. df['ClaimAdminId'].isnull())|(df['VisitType'].isnull())</td><td>Missing values in the ClaimAdminId, VisiType fields</td></tr>
<tr><td>6. (df ['DateFrom']<datetime.datetime(2016, 5, 15))|\
(df ['DOB']>df_items['DateFrom'])|\
(df ['DateClaimed']<datetime.datetime(2016, 5, 15))|\
(df ['DateClaimed']<df['DateFrom'])</td><td>Incoherence in the date related fields</td></tr>
<tr><td>7. df['HFID']!=df['HFId']</td><td>Check if ClaimAdminID has the same HFID as the ClaimHFID</td></tr>
</table>

Aggregation

Field name	Description
LastSameItem	Number of days since last submitted (and accepted) item (same ItemID)
SameItemPerDay/Claim	Count of items having the same ItemID, submitted same day/claim
ItemPerClaim AmountPerClaim/Day	Count of items having the same ItemID, submitted within the claim Amount related to the claim/day (ItemPrice or PriceAsked?)
ItemsPerWeek/Month/ Quarter/Year	Count of items for same Insuree over a period of 7 days (1 week), 30 days (1 month) or 90 days prior to the current submission
AmountPerWeek/Month/ Quarter/Year	Total ItemPrice/PriceAsked for the items submitted over a period of 7, 30 or 90 days prior to the current submission
AverageClaimOverMonth	Average amount of a claim over a 30 days period until the current submission AmountPerMonth/ClaimsPerMonth
AverageOverQuarter	Average amount / week related to claims submitted over past 90 days AmountPerQuarter/12
AverageDailyOverMonth	Average amount/day related to claims submitted over past 30 days AmountPerMonth/30
IsPackage	Check is a package was submitted within the associated claim

!Infinite possibilities of aggregation with respect to other Features and Time periods.

Challenges on data analysis



Only 3.78% of labeled data is rejected, while only 2.29% has an associated rejection reason

- Most of the features are categorical
  - Numerical: QuantityProvided, PriceAsked, ItemPrice Duration, DurationClaimed, Age, LastSameItem, SameItemPerClaim, SameItemPerDay, ItemsPerDay/Week/Month/Quarter/Year, AmountPerDay/Week/Month/Quarter/Year
  - Categorical:
    - Date: ~~DateFrom, DateTo, DateClaimed, DOB~~
    - Related to categories: ItemFrequency, ItemPatCat, ItemLevel, VisitType, HFLevel, HFCareType, Gender, ~~ItemServiceType~~, Poverty
    - ID related: ItemID, ~~ClaimID, ClaimAdminID, HFID, LocationID, HFLocationID, InsureeID, FamilyID~~, ICDID, ICDID1

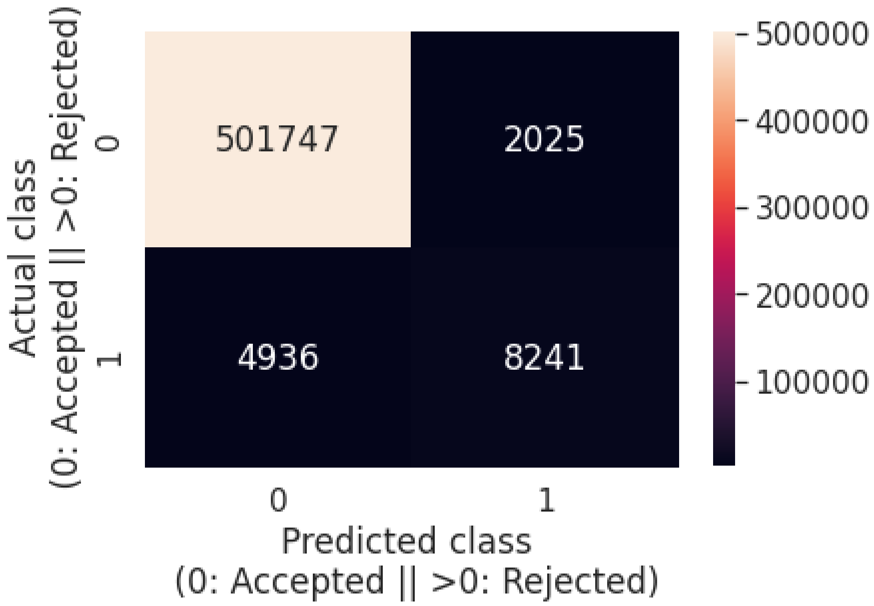
      or UUID/Code related: ItemUUID, ~~ClaimUUID, ClaimAdminUUID, HFUUID, LocationID, HFLocationID, InsureeUUID, FamilyUUID~~, ICDCode, ICD1Code

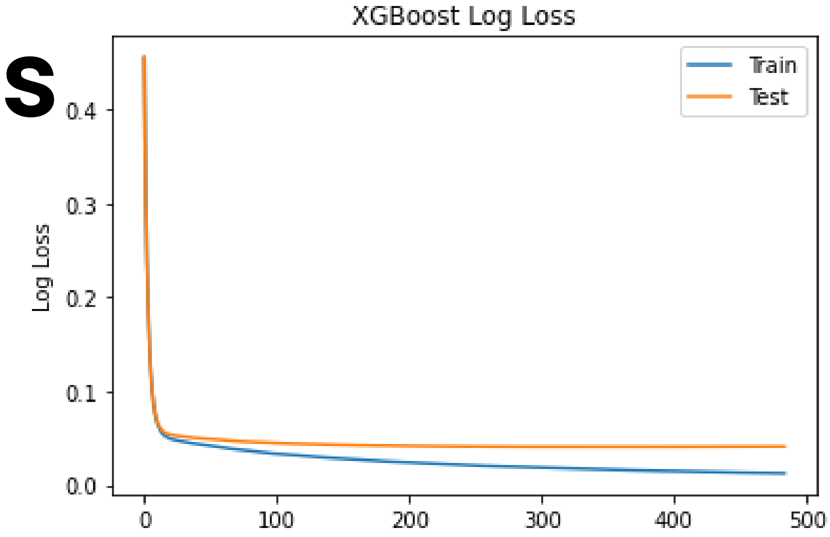
Aggregation – in practice

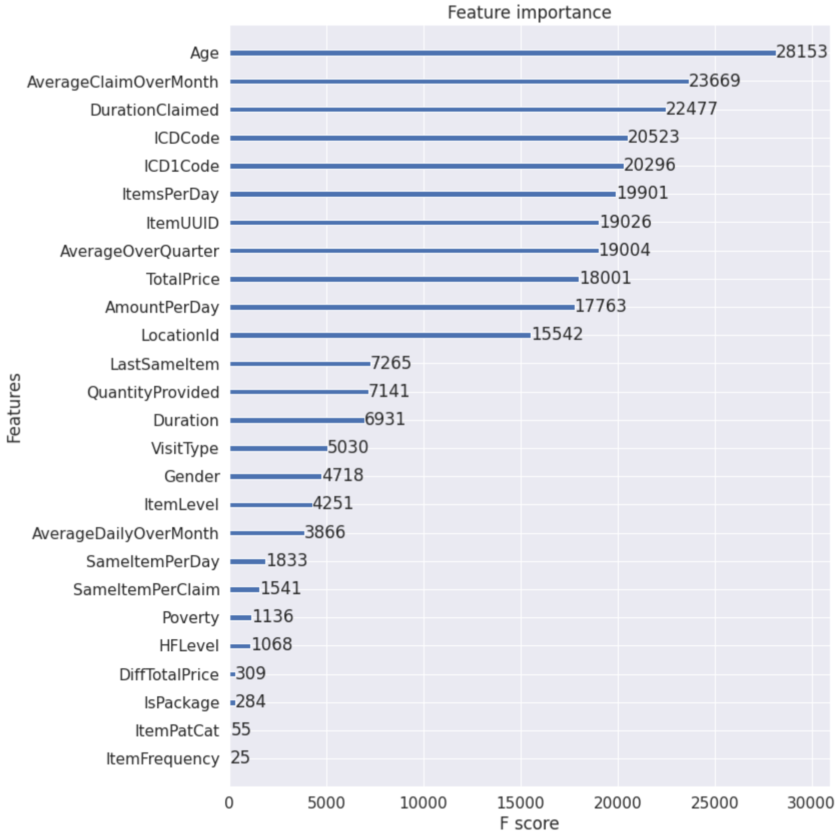
- In order to create the aggregated features, access to historical dataset is necessary

- For new submitted claims, in order to create the aggregated features for these claims, we need to retrieve the historical claims related to the InsureeIDs

AI model performances







$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} = \frac{501747+8241}{501747+2025+4936+8241} = \frac{509988}{516949} = 0.9865$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{8241}{2025+8241} = \frac{8241}{10266} = 0.8027$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{8241}{8241+4936} = \frac{8241}{13177} = 0.6254$$

$$\text{F1} - \text{score} = 2\,\frac{Precision * Recall}{Precision + Recall} = 2\,\frac{0.8027 * 0.6254}{0.8027+0.6254} = 0.7031$$

openIMIS

AI model performances

Dataset	Accuracy	Precision	Recall	F1-Score
Training set	0.9955	0.9600	0.8612	0.9079
Test set	0.9865	0.8027	0.6254	0.7031
Production: 1 weeks	0.9854	0.6424	0.4763	0.5471
Production: 1 month	0.9839	0.6340	0.4358	0.5166
Production: 2 months	0.9832	0.6156	0.4143	0.4953
Production: 3 months	0.9831	0.6028	0.4027	0.4828
Production: >1year	0.9853	0.3876	0.3167	0.3486

Fairness indices

- We can check the fairness of the AI model with respect to several feature values:
  Gender, Poverty, Age, Location, Race, Education, Religion, …

openIMIS

Top view of developments/tests

Classification based algorithms: **Decision Trees**, **Random Forests,** Isolation Forest, Bayesian Networks, ...

Nearest-Neighbor based algorithms: **k-NN**, Local Outlier Factor (LOF), Connectivity-based Outlier Factor (COF), ...

Clustering based algorithms: K-means, Cluster based Local Outlier Factor (CBLOF), Local Density Cluster based Outlier Factor (LDCOF), ...

Statistics based techniques: Parametric techniques, Non-parametric techniques, ...

Neural Networks related techniques: Artificial Neural Networks, Autoencoders, Long Short Term Memory (LSTM), ...

Top view of developments/tests

Classification based algorithms: Decision Trees, Random Forests, Isolation Forest, Bayesian Networks, ...

Nearest-Neighbor based algorithms: k-NN, Local Outlier Factor (LOF), Connectivity-based Outlier Factor (COF), ...

Clustering based algorithms: K-means, Cluster based Local Outlier Factor (CBLOF), Local Density Cluster based Outlier Factor (LDCOF), ...

Statistics based techniques: Parametric techniques, Non-parametric techniques, ...

Neural Networks related techniques: Artificial Neural Networks, Autoencoders, Long Short Term Memory (LSTM), ...

ML algorithm dependencies and variations:

- **Dataset variations**:
    - Binary class
    - Multiclass

    - Imbalanced case
    - Balanced case: Undersampling/Oversampling techniques (only on the training set)

    - Feature aggregation

- **Splitting of the dataset** in several sets: train/dev/test set, train/test set, ... (depending on the ML algorithm, validation method)

- **Hyperparameters** of the ML algorithm to be tuned

- **Evaluation metrics**: precision, recall, f1 score, accuracy, ...

- **Validation step:** holdout method, cross-validation, ...

Now what?

- Creation of a synthetic dataset that can be used for a DemoServer
- Create a video presenting the model?
- How to increase acceptance of the openIMIS AI module?
- How to improve the AI model?

Now what?

- Creation of a synthetic dataset that can be used for a DemoServer
- Create a video presenting the model?
- How to improve acceptance of the openIMIS AI module?
- How to improve the AI model?
  - Still needs to be improved?

openIMIS

Thank you

Contacts SwissTPH:

- Dragos Dobre (dragos.dobre@swisstph.ch)
- Simona Dobre (simona@dobre.fr)
- Siddharth Srivastava (siddharth.srivastava@swisstph.ch)

More information on openIMIS

Website: www.openIMIS.org

Wiki: wiki.openIMIS.org

Source code: github.com/openimis

Documentation: docs.openIMIS.org

Demo: demo.openIMIS.org

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung

giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

digital square

Swiss TPH
Swiss Tropical and Public Health Institute

DEVAIS

SolDevelo