AUTOMATIC IDENTIFICATION OF UNPLANNED HOSPITAL RE-ADMISSIONS

DATASHEET

Motivation

- For what purpose was the dataset created?

The original dataset was created with the purpose of identifying those readmissions to any of my organisation's hospitals that happened within 31 days from the previous discharge, and that were unplanned and concerning a complication related to their original treatment. In order to use this dataset for this project, I augmented it with additional information / columns coming from different data feeds from the EHR system used at my organisation.

- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset?

The dataset was created by my team with the manual classification of each case (as planned or unplanned) completed by the Governance teams in my organisation. The creation of the dataset was originally requested by my department, and "funded" by my organisation.

Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances comprising the dataset represent single hospital re-admissions and include details on the re-admission episode, alongside details of the previous admission, as well the details of when the episode was classified and by who.

- How many instances of each type are there?

There are 50'608 manually classified instances in the original dataset, before any filters was applied. The original dataset contained a 93%-7% split of planned (47'392) vs unplanned (3'216) cases.

- Is there any missing data?

In the augmented dataset 8'178 cases had missing information, therefore they had been removed from the analysis.

- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals’ non-public communications)?

The dataset contains confidential information about the patients treated, their episode of care, and the medical team that treated the patients, as well as the person that completed the classification of the case.

For this reason, after discussion with the learning facilitator, the dataset will not be made available on the repository. In its place a dummy dataset will be discussed in the attached Jupyter Notebook, accompanied by an extensive description of the original dataset.

Collection process

- How was the data acquired?

The underlying data is/was generated by a script that runs once a day and parses one of tables coming from the EHR system used in my organisation. This table, the “admissions” table, contains a record of all visits and admissions of patients to any of the hospitals within my organisation’s empire. Specifically, the script looks for any new admission of patients that have been discharged from the same hospital within 31 days prior. For any episode found by the script, a new row is added to a SQL table to store the details of the 2 connected hospital admissions. This table is then made available to the governance teams in my organisation via an App built internally. Through the user-interface of the App, the relevant governance facilitator can review each case and assign a classification as to whether the re-admission was planned or unplanned. The classification is therefore safely stored in the table, alongside log information about who classified the case and when.

- If the data is a sample of a larger subset, what was the sampling strategy?

The data used in this reasearch is an augmented version of the data collected for the purpose described above. The augmentation (in terms of additional features) was performed by joining the original dataset to several other data feeds coming for the EHR system used in my organisation.

The final dataset used in this analysis, as discussed in the Jupyter notebook, is a subset in terms of rows of the original dataset, in order to:

- achieve parity between the 2 classes I was trying to predict (planned or unplanned);
- remove rows with missing information.

- Over what time frame was the data collected?

The data was collected between July-2017 and October-2024.

Preprocessing/cleaning/labelling

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

The pre-processing steps consist of:

- removal of rows with empty fields
- balancing the occurrences between the 2 classes
- Engineered features:
  - age at admission
  - boolean to indicate whether the new admitting consultant was the same as the admitting consultant in the previous episode of care
  - boolean to indicate whether the new attending consultant was the same as the attending consultant in the previous episode of care
  - boolean to indicate whether the service line of the new admission was the same as the service line of the previous admission (e.g. cardiac and cardiac)
  - boolean to indicate whether the new ward where the patient was admitted was the same ward of the previous admission
  - Day of the week of the new admission
  - Time of the day of the new admission
  - Day of the week of the previous admission
  - Time of the day of the previous admission
  - Discretisation of the payer type field into ‘self-pay’, ‘insurance’, ‘embassy’

- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data is saved and available only to my organisation.

Uses

- What other tasks could the dataset be used for?

The dataset can help identify common causes and factors influencing readmissions. It can also be used to stratify patients into different risk levels for readmission. This information can support medical staff in creating tailored follow-up plans, such as using phone or telemedicine for high-risk patients, or gathering additional clinical data and vitals to guide discharge decisions.

- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The dataset was collected from a single private healthcare provider in the UK, encompassing several hospitals within the same organisation. As a result, the cohort analysed is not representative of the general population, as private healthcare primarily serves middle- and upper-class individuals. Additionally, while the organisation offers care for numerous conditions and procedures, the dataset may not cover all clinical needs comprehensively.

The classification of readmissions as planned or unplanned was carried out by different individuals, but inter and intra validation analyses have not been performed. This means that the extent of the bias or subjectivity could not be calculated.

However, to mitigate the subjectivity, all unplanned cases underwent a rigorous double-validation process: first by the governance team at the respective hospital and then by the central governance team of the organisation. In cases of disagreement, both parties collaborated to reach a resolution. Additionally, a random selection of 5% of planned cases was subjected to double validation to ensure accuracy and smooth and subjectivity across sites.

- Are there tasks for which the dataset should not be used? If so, please provide a description.

This dataset should not be used to generalise classifications beyond the organisation, as variables such as the hospital of admission and the location of admission (e.g., ward) may introduce biases and prevent fair algorithmic classification.

Distribution

- How has the dataset already been distributed?

This dataset is not for distribution, but it forms the basis for a revised dataset that is shared with PHIN for regulatory reasons.

- Is it subject to any copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is under applicable terms of use. The script that generated the original worklist used as the basis for the manual classification is under intellectual property license.

Maintenance

- Who maintains the dataset?

The dataset is maintained by my team.

MODEL CARD

Model Date: 10/12/2024
Author: Learner #704

Model Details:

This model classifies hospital readmissions into planned or unplanned, where an unplanned readmission is where a patient, who has been previously treated in hospital, has to return to the same hospital as an emergency within 31 days of being discharged for a problem related to their original treatment.

The model also returns, for each classification, the associated probability score, providing a measure of confidence in the model's decision.

In order to make the classification, the model needs a list of input parameters, specified in the input list below.

This automatic classification is meant to be used by the hospitals governance teams to identify the unplanned cases for:

- Internal reporting of the associated KPI
- internal learnings and improvement initiatives,
- external reporting to regulatory bodies such as the Care Quality Commission (CQC) or PHIN.

Model type:

Logistic Regression

Model Specifications:

```
LogisticRegression(penalty = 'l2', C = 0.2, solver='lbfgs')
```

Inputs:

- 'FacilityID': Hospital ID
- 'LocationID': Ward ID of the new admission
- 'InpatientServiceID': ServiceLine (e.g. cardiac, oncology ...) of the new admission
- 'AdmissionPriority': Priority of the admission
- 'Days_of_Diff_PrevDis': No. of days between previous discharge and new admission
- 'Prev_LocationID': Ward ID of the previous admission
- 'Prev_ServiceID': ServiceLine (e.g. cardiac, oncology ...) of the previous admission
- 'Sex': Patient's biological gender: F if female, M if male
- 'FinancialClassID': Payer type (e.g. self-pay, insurance, embassy ...)
- 'RaceID': Patient's ethnicity
- 'los_days': No. of days spent in the hospital in the new episode
- 'Prev_los_days': No. of days spent in the hospital in the previous episode
- 'n_thvis_new': No. of theatre visits in the new admission (i.e. is this a surgical patient)
- 'Prev_n_thvis': No. of theatre visits in the previous admission

- 'age': Patient's age at admission (new)
- 'has_same_admit_cons': TRUE is the admitting consultant of the new admission is the same of the previous one;
  FALSE otherwise
- 'has_same_attend_cons': TRUE is the attending consultant of the new admission is the same of the previous one;
  FALSE otherwise
- 'dow_prev_disch': Day of the week of the previous discharge
- 'time_prev_disch': Time of the day of the previous discharge
- 'dow_new_adm': Day of the week of the new admission
- 'time_new_adm': Time of the day of the new discharge
- 'has_same_service': TRUE is the service line of the new admission is the same of the previous one;
  FALSE otherwise
- 'has_same_location': TRUE is the location (ward) of the new admission is the same of the previous one;
  FALSE otherwise

Output:

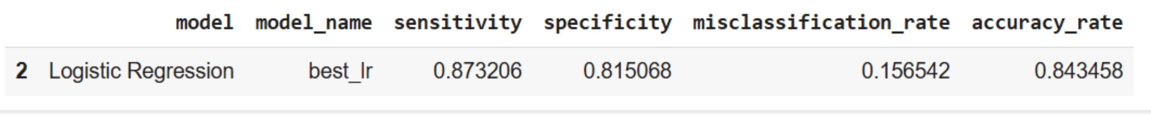
- Boolean - answering the question whether the re-admission was unplanned (True) or planned (False)

Model Performance Measures:

- Sensitivity
- Specificity
- Misclassification Rate
- Accuracy

The final choice was made by giving more importance to sensitivity and misclassification rate.

Performance of the logistic regression model analysed on a test dataset not used during the training phase of the model. The Test dataset consisted of 856 cases, with a split of 438 False (51.16%) (Planned) and 418 True(Unplanned) (48.83%).



	model	model_name	sensitivity	specificity	misclassification_rate	accuracy_rate
2	Logistic Regression	best_lr	0.873206	0.815068	0.156542	0.843458

Model Intended Use:

The model is designed to enhance the efficiency of governance workflows (only in my organisation) by embedding automated classification into an existing tool while maintaining a human-in-the-loop approach for trust, validation, and continuous learning. Key features of the intended use include:

- Integration into the App: Automates routine classification tasks, enabling governance teams to focus on more ambiguous cases.
- Confidence Transparency: Provides classification results along with confidence scores (e.g., probabilities in Logistic Regression) to help teams assess the reliability of predictions.
- Risk Management: Highlights edge cases with lower confidence for manual review, mitigating risks associated with misclassification and ensuring regulatory compliance.
- Workflow Streamlining: Facilitates cross-referencing of automated results with internal notes, maintaining oversight and improving process efficiency.

Model Limitations:

This model is to be used only within my organisation, as it wouldn't generalise well outside of it. It uses fields such as Location and Hospital, that are very specific to this organisation and have been included only because they can serve as proxies or indicators of the level of care complexity. Specific locations or facilities might correlate with the type of healthcare challenges encountered (e.g., specialised centers dealing with high-acuity patients).

Model Trade-offs

Potential trade-offs concern cases in which the model is to be used:

- at a new location, not yet synonym of complexity, because those instances won't be represented in the training dataset yet.

- at a location that has recently changed they type of patients and complexities treated.

Additional trade-offs are to be further explored once the next steps of the projects will be completed. At that point I will add all cases where the model has performance issues.

HYPERPARAMETER OPTIMISATION

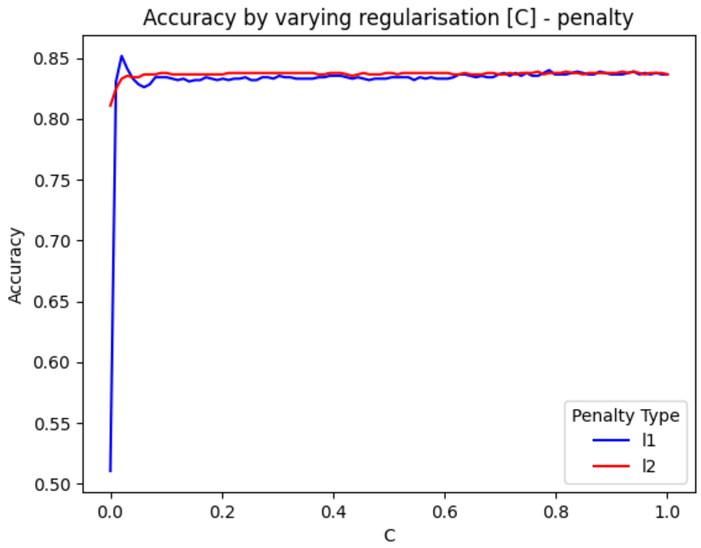
The chosen model was a Logistic Regression Binary Classifier.

The hyperparameters that have been optimised in this case are:

- **c**: Regularisation strength; smaller values specify stronger regularisation.
- **penalty**: The type of regularisation to apply ('l2', 'l1', 'elasticnet', or None)

To optimise these hyperparameters, a grid-search approach was implemented. The space searched is represented below, together with a plot showing the results obtained in terms of Accuracy on the validation dataset.

```
# Define parameter grid
step_c = 100
c_values = np.linspace(10e-5, 10e-1, step_c)
penalty_types = ['l1', 'l2']
```



This optimisation was performed based on accuracy, and was run on a validation dataset (856 cases) that was not used during the training phase of the model.

RESULTS

Four different models were explored and their results have been compared across the selected performance metrics:

- On a test data set that had not been used to train the models
- After each model underwent hyperparameters tuning (for reference, the tuned models have been reported below)

DECISION TREEE

```
best_clf = tree.DecisionTreeClassifier(
        criterion = 'gini',
        splitter = 'best',
        max_depth = bestdepth,
        min_samples_split = 2,
        min_samples_leaf = 1,
        max_features = None,
        random_state = None
        )
```

RANDOM FOREST

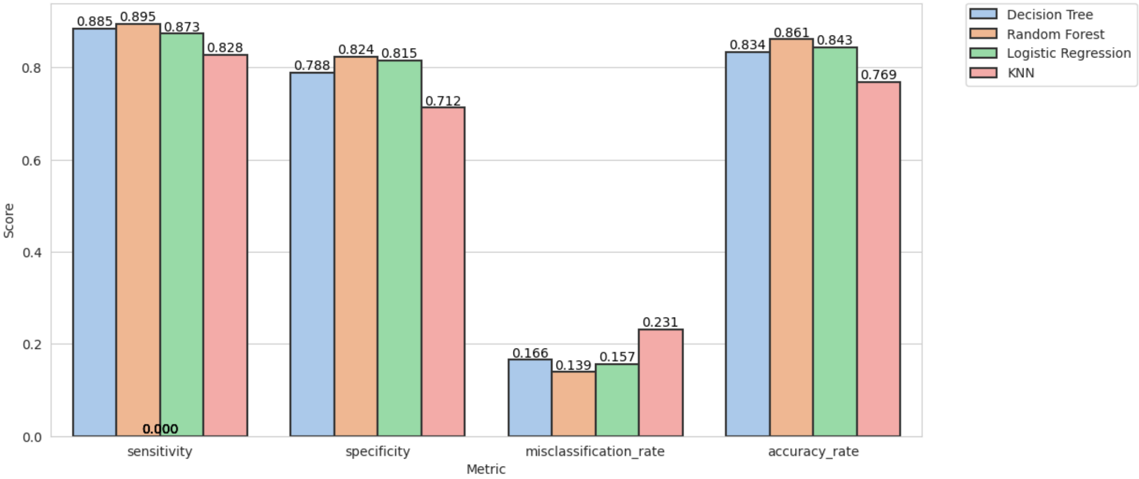
```
best_rf_clf = ensemble.RandomForestClassifier(n_estimators=150,
max_features="sqrt", bootstrap=True)
```

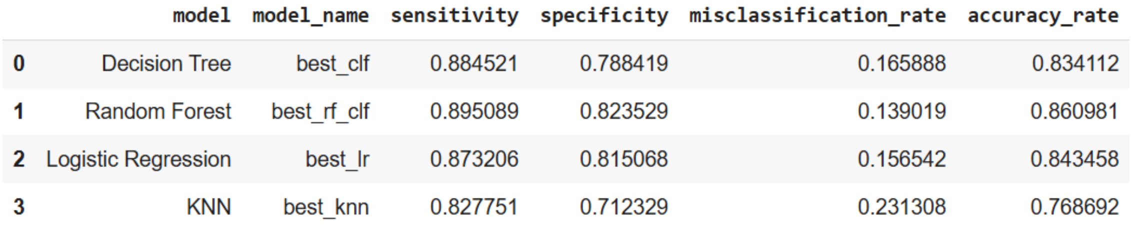
LOGISTIC REGRESSION (model chosen for this application)

```
best_lr = LogisticRegression(penalty = 'l2', C = 0.2,
solver='lbfgs')
```

KNN

```
best_knn = KNeighborsClassifier(n_neighbors = 7)
```





	model	model_name	sensitivity	specificity	misclassification_rate	accuracy_rate
0	Decision Tree	best_clf	0.884521	0.788419	0.165888	0.834112
1	Random Forest	best_rf_clf	0.895089	0.823529	0.139019	0.860981
2	Logistic Regression	best_lr	0.873206	0.815068	0.156542	0.843458
3	KNN	best_knn	0.827751	0.712329	0.231308	0.768692

Random Forest (best_rf_clf) has the highest sensitivity of 0.895 indicating it is the best at detecting positive cases. This model balances well with a high specificity (0.824) and a relatively low misclassification rate (0.139), leading to a high accuracy rate of 0.861.

Decision Tree (clf) achieves a sensitivity of 0.885, which is slightly lower than Random Forest but still strong. Its specificity (0.788) and accuracy (0.834) are lower than those of Random Forest, but it maintains a reasonable balance.

Logistic Regression (best_lr) has a sensitivity of 0.873, placing it third in terms of positive case detection. Its specificity (0.815) and accuracy (0.843) are comparable to the Decision Tree but slightly behind the Random Forest.

KNN (best_knn) shows the lowest sensitivity of 0.828, indicating it struggles the most with identifying positive cases compared to the other models. Its specificity (0.712) and accuracy (0.769) are also the lowest among all models.

The choice of which model to pick was done on the basis of performance, explainability and deployment.

The logistic regression model was chosen because:

- Comparable high performance to the best performing model (random forest)
- High explainability and transparency
- Added bonus of the probability score associated with each classification, providing a measure of confidence in the model's decision