LISBON DATA SCIENCE ACADEMY

Reducing wrongful discharge

Predicting wrongful discharge at the Hazel and Bazel Hospital (Report II)

Prepared for:
The Hazel and Bazel Hospital

Prepared by:
Yash Pandya
Lisbon Data Science's *Awkward Problem Solutions™*

{Day} of {Month} 2022

Table Of Contents

1. Business Conclusions

1.1 Summary

The performance and requirements provided in the initial exchanges with the client and highlighted in report 1 are as follows:

1. To have sufficiently high recall so as to effectively minimize wrongful discharges (those leading to readmission within 30 days).
2. To have at least 50% precision such that at least half of patients predicted for readmission are actually sick.
3. To reduce discrepancies between readmission rates of subgroups (gender, age, race and insurance status and admission source) to less than 10%, and differences between medical specialties to less than 5%.

The first testing phase consisted of slightly less than 10,000 total prediction requests being sent to the /predict endpoint of the API (see section 2.3 for analysis of this data). Of these, the deployed model generated and stored 9667 predictions (6174 False and 3493 True) (Fig 1 A). Subsequently, the true values, indicating whether patients were actually readmitted or not were provided using the /update endpoint (8570 False and 1097 True) (Fig 1 B). This data was used to calculate the various performance metrics (precision, recall, F1-score) required to assess model performance.

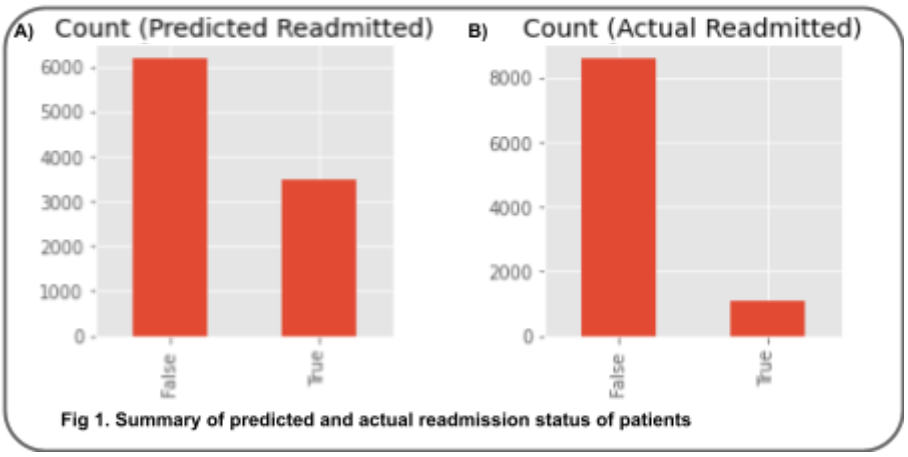


Fig 1. Summary of predicted and actual readmission status of patients

As shown by this data, the deployed model predicts a large amount of patients will require readmission when they do not. The predictions have a precision score of 17.3%, which is lower than the 50% target which was initially set. However, the model still correctly identifies over half of the patients who do require readmission (recall score of 55%), indicating that it is achieving the primary goal of reducing readmission rates.

When examining the specified sensitive features for signs of potential discrimination, the model does meets the requirements for gender (difference between male and female precision rates <1%), insurance status (difference between insured and uninsured patient precision rates of 2%) and admission sources (maximum difference of 7%). However, there is a large discrepancy when looking at patient race (maximum difference of 16%) which appears to stem from discrimination against asian patients (28%, whereas all other groups range from 12 to 17%). The model also maintains signs of discrimination against young adults (max difference of 23%, with 20-30 and 30-40 Age groups being most likely to be readmitted). Lastly, the model does not reduce differences between different medical specialties to within 5% (maximum score of 31%). These results suggest that it is important to address potential flaws in the current model, but also to potentially redefine the targets and account for inherent differences in the needs of different populations, and differences in the nature of illness treated by various medical specialties.

2. Results Analysis

2.1 Model Performance

The model was designed and trained with the use of a dataset provided by the client including over 80,000 individual admissions. This data was unbalanced (11% of patients were readmitted, and the remainder were not). This informed the selection of a Random Forest Classifier, trained with random oversampling of the true set to generate a balanced dataset. That approach was selected to minimize overfitting to the training data (helping to keep the model robust for the test sets), and to minimize the impact of this unbalanced dataset.

The metric used to assess the model during the development were the F1 score, precision, recall and ROC-AUC scores. It should be noted that while accuracy was calculated and reported (accuracy score of just under 70%), this metric was not selected for evaluation or optimization. This is because the dataset is not balanced, allowing the model to achieve a high accuracy by predicting that no patients would be readmitted. Additionally, in this case, returning a False negative is significantly more costly (to the patient and, potentially, to the hospital), as such, the F1 score is a more suitable metric for this specific problem.

During development, the provided data was separated into a train set and a validation set (90% and 10% of entries respectively). The 'expected' scores for the various metrics shown are based on predicted values of the validation set. During the first testing phase, slightly less than 10,000 requests were made to the deployed API /predict endpoint, followed by a similar number to the /update endpoint. The model generated 9667 predictions for the admissions provided and stored the true result for these values as well (Fig 1, section 1). The 'actual' scores for the various metrics are based on these predictions and true values.

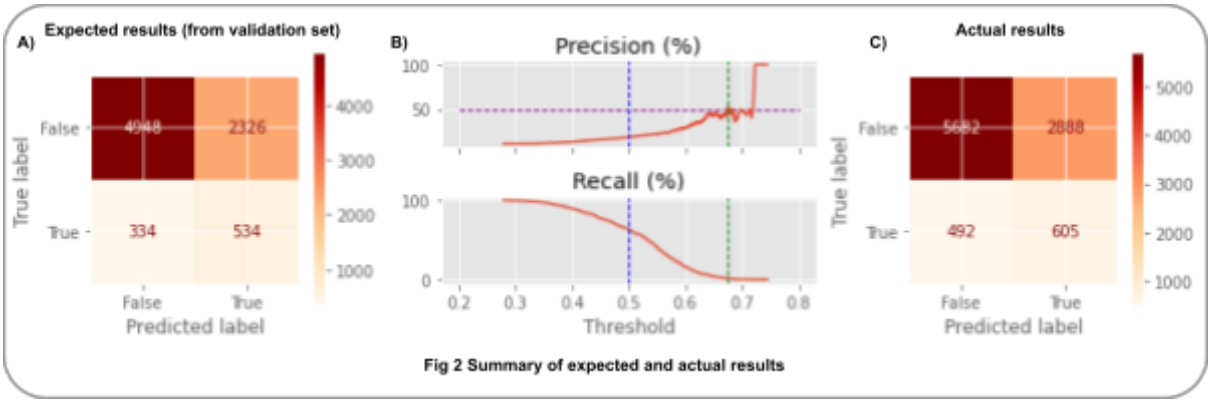


Fig 2 Summary of expected and actual results

A confusion matrix generated for predictions based on the validation data (Fig 2A) reveals that the model is expected to correctly identify more than half of the patients who would be readmitted. This data also reveals that approximately 80% of the positive results will be false positives (Fig 2 A). Precision and Recall based on the validation dataset are illustrated (Fig 2B), showing that there is a large tradeoff between the two metrics. In particular, it is important to note that increasing the threshold from the employed one of 0.5 (indicated by the dashed blue line) would not provide large increases in precision without a massive reduction in recall (for details on threshold selection, see report 1). Following from this, if aiming for 50% precision (purple line, one of the technical requirements explored in section 2.2), an extremely high threshold has to be selected (green line), resulting in very low recall. A confusion matrix generated for the predictions and true values provided during the test (Fig 2C) shows very similar behavior to the validation set, although it has slightly lower performance on the key metrics employed (Table 1).

Table 1 provides the three key metrics (accuracy, precision and the F1 score), a practical definition of what this means with respect to the aims, the scores of the model when examining the validation set (expected scores) and the scores when assessing the predictions made (actual scores). The results showed a lower performance than expected with respect to the three key metrics selected to evaluate performance. Specifically, the recall score was 6% lower (actual score of 55% vs expected score of 61%), and the precision score was 2% lower (actual score of 19% vs expected score of 17%), resulting in an F1 score that was slightly lower than expected. ROC-AUC curves were also generated for the expected and actual performance of the model and also showed a slight decrease in model performance (Annex.1 Figure 1).

Metric	Definition	Expected Score	Actual Score
Recall score	From all patients who will be readmitted, the proportion of patients that the model correctly identified as being readmitted.	61%	55%
Precision score	From the number of patients identified by the model as being readmitted (positive predictions), the proportion of patients who actually were readmitted.	19%	17%
F1 score	A weighted average of precision and recall. Optimizing for this metric allows for a balance between precision and recall	28.9%	26.4%

Table 1. Expected and actual model performance

This discrepancy in results may be linked to several factors, all of which will require further investigation:

1. It is possible that the deployed model was overfit to the training dataset which was originally provided. This can be addressed by potentially considering alternative train/validation splits and by revisiting the hyperparameter tuning step
2. The training data may not be representative of the test data. Considering changes in patient care and staff over time, it is possible that the training data is not an accurate reflection of the current state of hospital admissions. The new data will be analyzed to determine if this is indeed the case (see section 2.3)

In summary, the actual performance of the model was slightly lower than expected based on the scores generated with the training data.

2.2 Success on requirements

The requirements provided by the Hazel and Bazel Hospital in the initial exchanges with Dr Agnes Crumplebottom, and as a result of clarifications requested are the following:

1. The primary aim is to minimize wrongful discharge, as such the model should have sufficiently high recall so as to correctly identify a large proportion of patients who would be readmitted.
2. It was further emphasized that the model should have at least 50% precision such that at least half of patients predicted for readmission are actually sick. The importance of this measure was stressed from a financial and practical point of view, as the hospital cannot keep patients admitted indefinitely.
3. Lastly, there is a suspicion that there may be discrimination occuring at the hospital, and that, as a result, patients with specific demographic characteristics may be receiving poorer care and be more likely to be wrongfully discharged. Consequently, an additional requirement is to mitigate the impact of this problem and reduce discrepancies between readmission rates of subgroups (gender, age, race and insurance status and admission source) to less than 10%, and differences between medical specialties to less than 5%.

During initial development, a threshold which would allow 50% precision was selected (Fig 2B, purple line). This value was quite high (Fig 2B, green line, threshold of 0.67). Using this threshold meant that a tiny fraction of readmitted patients would be identified (very low recall values) and the primary objective of minimizing wrongful readmission would not be achieved. Consequently, and as explained previously, a threshold of 0.5 was employed, reducing the precision of the model and increasing the recall, allowing for a majority of patients who would be readmitted to be correctly identified.

As a result, the primary aim of minimizing wrongful discharge was achieved, with 55% of sick patients being correctly identified (recall of 55%). However, the requirement for at least 50% precision was not satisfied (actual precision of 17%). This model results in a large number of patients being flagged as potential wrongful discharges when, in fact, they could have been released from the hospital. While this is a safer method with respect to patient outcomes and legal liability of the hospital, it may have a financial impact, and will influence resource availability. Consequently, this is something which should be improved upon if possible.

When addressing potential discrimination, it was required that differences between gender, age, race, insurance status and admission source be limited to less than 10%, and differences between medical specialties be limited to less than 5% in the predictions provided of the model. To assess this, precision scores for each of the subgroups were compared for different subgroups with at least 50 representatives in the test data set. The maximum difference of the precision for subgroups within each feature was used as an indicator for discrimination, according with the targets provided. Results for this assessment are provided in Table 2. Expected results are those which were obtained based on the validation set during development and actual results are those obtained during the testing phase. Categories highlighted in Green are those for which technical requirements were met, and those in red are those where they were not met.

As shown on the table, the difference in precision for predictions made in male and female patients is less than 1% (the value shown is rounded down to 0.00), well below the target of 10%. A similar result is observed for patient insurance status and patient admission sources, where the maximum difference between groups based on these sensitive features is below the target value of 10%. In the case of admission sources, it should be noted that the difference in precision, while still below the target of 10%, is much higher than the expected value (expected of 2%, actual of 7%). For these three characteristics, therefore, the discrepancies comply with the technical requirements.

Feature	Expected maximum difference in precision	Actual maximum difference in precision	Target maximum difference
Gender	1%	>1%	10%
Race	10%	16%	10%
Age	37.5%	23%	10%
Insurance Status	4%	4%	10%
Admission Source	2%	7%	10%
Medical Specialty	29%	31%	5%

Table 2. Compliance with technical requirements linked to discrimination

When assessing race, the expected maximum difference was 10% (the expected value was compliant with the requirements) whereas in the test set it was 16%, hence non-compliant. This value was directly linked to a disproportionate predicted readmission rate for asian patients who, along with hispanic patients, have higher rates of predicted readmission, although they have low representation in the data. This difference was due to an increase in the readmission rate observed in Asian patients, suggesting that the model may be discriminating against this group, a problem which needs to be further examined and addressed.

It was also found that there was a large value (23%) for the maximum difference when considering patients from different age groups failing to fall below the target of 10%. The source of this result was increased readmission rates in the 20-30 age group, which was also observed during analysis of the training data. The model thus did not comply with the technical requirement for age, and this requires further investigation.

Additionally, the model does not comply with the requirement for a minimum 5% difference between medical specialties, (actual value of 31%). Importantly, it should be noted that some specialties which were expected to have higher readmission rates (primarily pulmonology) were not represented in sufficient numbers in the test set. This value is not extremely concerning as different medical specialties will treat different kinds of diseases, and will therefore have different treatment targets and outcomes.

In conclusion, the primary objective of minimizing wrongful discharge (those leading to readmission within 30 days) was achieved. However, the high recall required to achieve this goal meant that the requirement for minimum precision of 50% was not achieved. With respect to minimizing discrimination, the model achieved the targets for gender, admission source and insurance status, but did not achieve the targets set with respect to race, medical specialty visited and age of the patient.

2.3 Population Analysis

Data drift (changes in the data over time), such as alterations in the demographics of patients coming to the hospital, or changes in the quality of care due to changes in methods, staff, or facilities for example may result in impaired model performance. To determine if the new data differs from the training set which was used for development of the deployed model, the various features and the readmission rate between the two datasets (referred to as training data and moment 1 data) were compared.

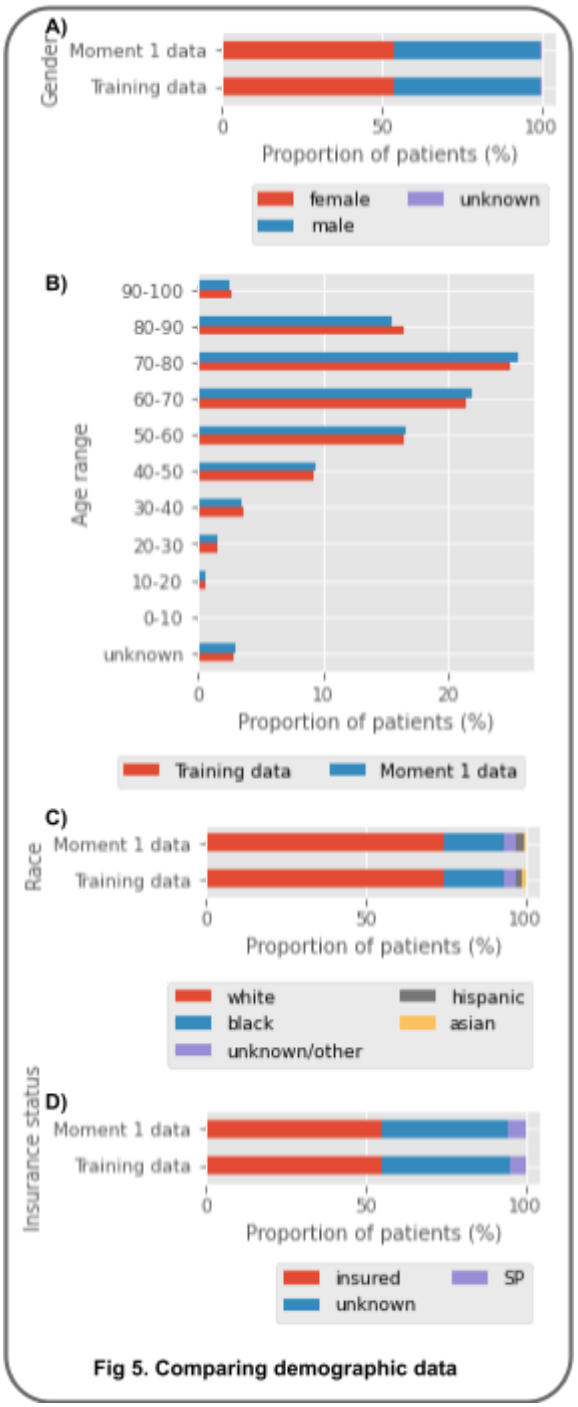


Fig 5. Comparing demographic data

Several requests which were made to the API were not stored, primarily those which did not contain a patient id or an admission id, consequently 9667 records were stored for the new dataset. The data, previously stored in JSON format was extracted and the categorical values were encoded as for the modeling stage, using the same pipeline components. This method ensures that the data is analysed in a way which allows us to determine if the data used to generate predictions was vastly different from the data used to train the model.

Initially, demographic data (gender, age, race) and patient insurance status were compared, to determine if any of these specific groups had an altered representation in the new dataset. As shown on Fig 5A, the proportion of male and female patients visiting the hospital was unchanged, with female patients composing slightly over 50% in both datasets (53% in training data and 54% in the data from moment 1), with a negligible number of patients for whom this information was unavailable, and the remainder being male. It is also apparent that there is no alteration in the age distribution of the patients (Fig 5B), as the age of patients tends to skew towards the elderly in both datasets. Furthermore, there is no difference in the racial make up of the patients, as white patients continue to constitute the majority (slightly under 75% in both datasets), black patients are a well represented minority (slightly under 19% in both datasets), while asian and hispanic patients have low representation (approximately 2% and 0.6% respectively in both datasets), with the remaining patients being categorized as unknown/other. With respect to patient insurance status, a similar tendency was observed, without significant alterations in the proportion of insured patients (55.4% and 55.3% in the training data and moment 1 data respectively) or in the proportion of patients who were self payers ('SP', 4.9% and 5.1% in the two datasets).

Having concluded that the demographic of the patients did not change between the training data and the data received during the first round of predictions, the next step is to examine the reasons for the patients going to hospital, and the clinical data available. As shown on Fig 6 A and B, and as was

observed for the training data, the largest portion of patients in the new data arrive at the hospital for emergency care (53.0% in the new data vs 53.4% in the old), and the proportions across other admission types are also remarkable consistent (Fig 6A), as are the various admission sources (Fig 6B). Furthermore, the top 5 most commonly visited specialties were also the same (from most common to 5th most common: internal medicine, emergency/trauma, family/general practice, cardiology and general surgery). These specialties continue to make up the bulk of patients for whom the visited specialty is known. It should be stressed that a large portion of patients are still not assigned a specialty, something which should be addressed in the future.

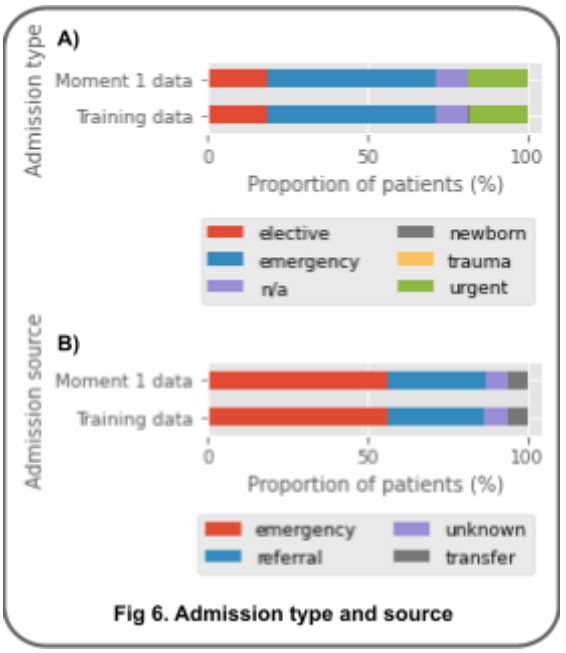


Fig 6. Admission type and source

The various numerical parameters pertaining to patient hospital stays such as the length of the stay, the number of visits during the last year (outpatient, inpatient and emergency visits), the number of medical interventions (procedures, medicines prescribed, laboratory procedures) were also briefly examined and showed a similar distribution between both sets of data. Additionally, the deployed model grouped diagnoses based on the 19 major categories based on CDC classification of diagnostic codes (parsed data available here). Examining the data after grouping reflects the consistency observed with the other features, as diagnoses associated with the circulatory system, the endocrine system, and the respiratory system continue to be the most common ones, while those linked to congenital disorders, pregnancy and injury/poisoning are among the least common.

Data linked to patient discharge was examined. This includes the discharge disposition code an indicator of where the patient will be discharged to (Fig 7A), which continued to show strong similarities between the training dataset and the newly received data.

Finally, considering the objective of predicting readmission within 30 days, it is important to ensure that there has not been a significant alteration in the proportion of patients who are being readmitted. Examination of the data verifies this to be the case (readmission rates are 11.1% and 11.3% in the training data and data from moment 1 respectively).

Based on this analysis, it can be concluded that there has not been a large change in the data that is being input into the model and that data drift has not been a significant contributor to changes in its performance.

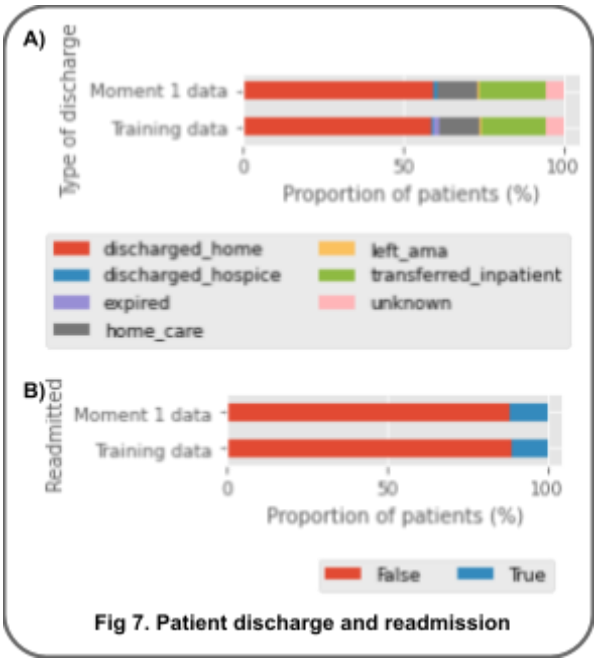


Fig 7. Patient discharge and readmission

3. Next Steps

3.1 Next Steps

Clearly, from the obtained and expected results, it is apparent that there is a need for improvements to be made to the performance of the model, allowing for the targets to be achieved and the requirements to be met.

In particular, among the alterations which could be made using the current dataset, it should be possible to make more effective use of the patient diagnoses. The deployed model grouped the diagnoses into several large categories, resulting in medical conditions which may have vastly different clinical presentations, outcomes and, consequently, a different potential for wrongful discharge, being combined into categories. This will have affected the ability of the model to distinguish these patients. This problem can be addressed by employing one hot encoding, such that each diagnostic code becomes a category (patients would get a 1 if they have that condition and a 0 if they do not). While this would allow for all of the information to be maintained, it would result in an extremely high dimensionality to the data (there are several hundred unique diagnostic codes in the dataset and a large number more which have yet to be used). This may prove to be problematic for rare conditions, and may result in impaired model performance beyond a certain point, so further approaches, such as limiting the number of categories to the top 100 most represented or text vectorization approaches may be warranted.

Additionally, the patient identification feature was not included, although it may have improved model performance. While this decision was made to ensure that the model works equally for all patients and treats each admission as a unique event, it ignores patient history. Future approaches using this feature would indirectly allow for patient medical history, primarily their previous readmission history, to be included. Exploring this further, examining previous admissions to the hospital to determine if they have had recurrent visits for the same conditions, or to the same specialty may allow for the identification of medical conditions for which readmission within a 30 day period is more likely. These approaches may allow for the building of additional features such as an expected readmission linked to certain medical conditions (such as a requirement for physiotherapy after surgery) or medical specialties. Depending on how this affects patient care, this may or may not be a relevant feature for determining whether a patient has been wrongfully discharged, or examining the care provided, and this is something which needs to be carefully assessed.

Importantly it is important to work with the management at the Hazel and Bazel Hospital to ensure that the objectives align with their goals. As such, it may be beneficial to reframe the problem in a way which redefines “wrongful discharge”. In particular, it would be beneficial to determine if the readmission occurring within 30 days of discharge is directly linked to the discharge, under this revised definition a wrongful discharge would help to more accurately identify situations where the quality of care may come into question and the hospital faces potential recriminations. Additionally, it would be important to include additional data, such as the admission and discharge dates as it is possible that the day of the week (weekday vs weekend), or the time of year (season, holidays or not) may influence the quality of care provided, and this feature would also allow for changes in staff, facilities or training to be accounted for.

4. Deployment Issues

4.1 Re-deployment

The slight drop in performance observed between the training dataset and the first round of requests suggested that there may be a need for some additional investigations into the model. Additionally, updates in several dependency packages during the month of February made the deployed app more prone to crashing (highlighting the need for ongoing monitoring and maintenance). Both of these issues were addressed and the model was retrained with the inclusion of the new data prior to being redeployed.

As identified in section 3.1, it is highly likely that the grouping and encoding of diagnostic codes, while allowing for the inclusion of all possible codes, would result in a significant loss of information. Consequently, an attempt was made to use one-hot encoding to treat the diagnostic codes. Unfortunately, this was not a reasonable approach with the hardware available and so two other approaches were explored. The first was to replace the icd9 codes with their full description, followed by vectorization and PCA analysis, and the second was to group codes together based on the likelihood of patients with those codes being readmitted, generating 3 additional features (‘diag_1_risk’, ‘diag_2_risk’ and ‘diag_3_risk’) which were then combined into categories and ordinally encoded. Both approaches gave a mild improvement (improved the ROC-AUC for the new samples from 0.61 to 0.62) and the latter was selected for ease of implementation.

For the initial deployment, the model was erroneously not trained on all available data, so this was the next step which had to be carried out, generating new pickled files for use in the app. These were added, as were the updated preprocessors which allow for implementation of new features. Furthermore, minor alterations were made to the code of the custom transformers and utility files to allow for improved readability of the code. In addition, while the app still aims to return a prediction as long as the admission and patient ID fields are available in the request, a few minor modifications have been made to the tests which check for data validity, and these can be easily altered to be more stringent if necessary.

Lastly, during redeployment, it became apparent that the app was currently making use of outdated packages (further explored in the next section). As a result, it was necessary to update the requirements for the redeployment to ensure that the app remains stable. During redeployment, the associated database of the app was also reset.

4.2 Unexpected problems

Problems were identified during several stages of deployment and during the week of requests sent to the app. In particular, the problems encountered were related to validating incoming requests, returning appropriate errors where necessary, logging and storage of logs, data storage and out of date dependencies resulting in crashes. A brief description of the problems encountered and the solutions implemented is provided below.

1. During the initial testing phase, prediction requests containing duplicated ‘admission_id’ values were not being correctly processed and, while not being stored, they would not return an appropriate error or warning. This was rapidly addressed to ensure that duplicated values generate a warning for the user.
2. The previously deployed version of the app did not allow for duplicate admission ID values, assuming that these values would be considered invalid or incorrect. However, it is apparent that this may not be the case and this aspect may be problematic for 2 reasons. The first of

these is that it would store the first prediction, but a patient may have a prediction generated on the first day after admission, and, depending on the result, stay in the hospital, with another prediction generated on the second day. This second prediction would not be stored, so this data would be lost. The second problem is that, due to the way 'actual_readmitted' status was being stored, it was possible for predictions (for the same admission) to not correspond to the relevant prediction. Consequently, depending on requirements, the storage can be adjusted to allow duplicated admissions to overwrite previous values.

3. The Heroku platform used for the deployment allows for retrieval of the last 1500 logs, suggesting that troubleshooting further back than this would be challenging. To increase the level of detail in the logs, additional details were added to the output and logs were downloaded and stored every few hours. Unfortunately, the outputs replaced significant amounts of the additional messages with '...', resulting in the loss of significant information. Solutions to be explored include making use of the loguru library, or using Heroku log drains to store them externally.
4. During the current redeployment, several problems were identified. The first was that, due to updates of the python package 'itsdangerous' (used for json processing in this case), there was an unaccounted difference between the development and deployment environments, causing the app to crash repeatedly. This problem was solved by explicitly specifying the inclusion of an older, compatible version of the package (2.0.1) within the requirements, but this should be addressed in the future to ensure compatibility with updated packages. Similar issues may arise in the future with the category encoders employed (they receive a FutureWarning highlighting the use of deprecated syntax in some areas).
5. Lastly, although the app and model are prepared to handle unexpected and inaccurate data (the training data had severe inconsistencies, and a large amount of missing information, which were mirrored in the data used for the predictions), it is still extremely surprising to find patients with multiple records who have different blood types recorded at different admissions. Making use of the blood type does improve the predictions generated by the model, so the feature was maintained, however, this is something worth investigating from both our end, and from the clients side. This could be prevented by implementing stricter validation codes, however, as the expectation is to generate a prediction as long as an admission ID is provided, this has not been implemented.

4.3 Learnings and Future Improvements

In summary, in addition to the steps explored in section 3.1 above, which may help to improve the models accuracy, a few additional improvements can be made to the deployed API. The first of these would be to upgrade storage capacity, as the current deployment only allows for 10,000 entries, meaning there is potential for data loss if it is not regularly backed up.

Next, it is essential to develop and implement improved logging methodology, or allow for retrieval of more than the last 1500 entries to ensure that the app can be effectively monitored and issues can be identified and resolved more easily. These problems include, but are not limited to, ensuring that the app remains up to date with requirements and dependencies - particularly as some of these may be linked to security.

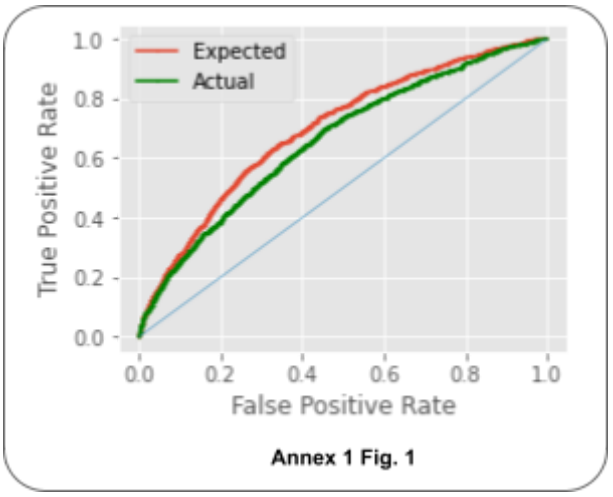
Finally, the training data was particularly inconsistent, with a large number of missing values and redundant definitions within some categories. While the model and initially deployed and current app are equipped to handle these problems, they come at a cost of model accuracy, and may potentially introduce security risks as the validation procedure for the data is extremely lax. Consequently it is extremely important to work, in collaboration with the team at the Hazel and Bazel Hospital, to ensure improved record keeping, and clearly define rules for data which should and should not be accepted by the mode. This would allow for more stringent rules to be implemented regarding data validity for

predictions and identify records which are clearly erroneous or do not have sufficient information to generate predictions.

5. Annexes

Annex 1

The ROC-AUC curve was also evaluated to assess expected and actual model performance (shown on Annex 1 Figure 1). This shows that the actual performance (green) was slightly worse than the expected performance (red)



Annex 1 Fig. 1