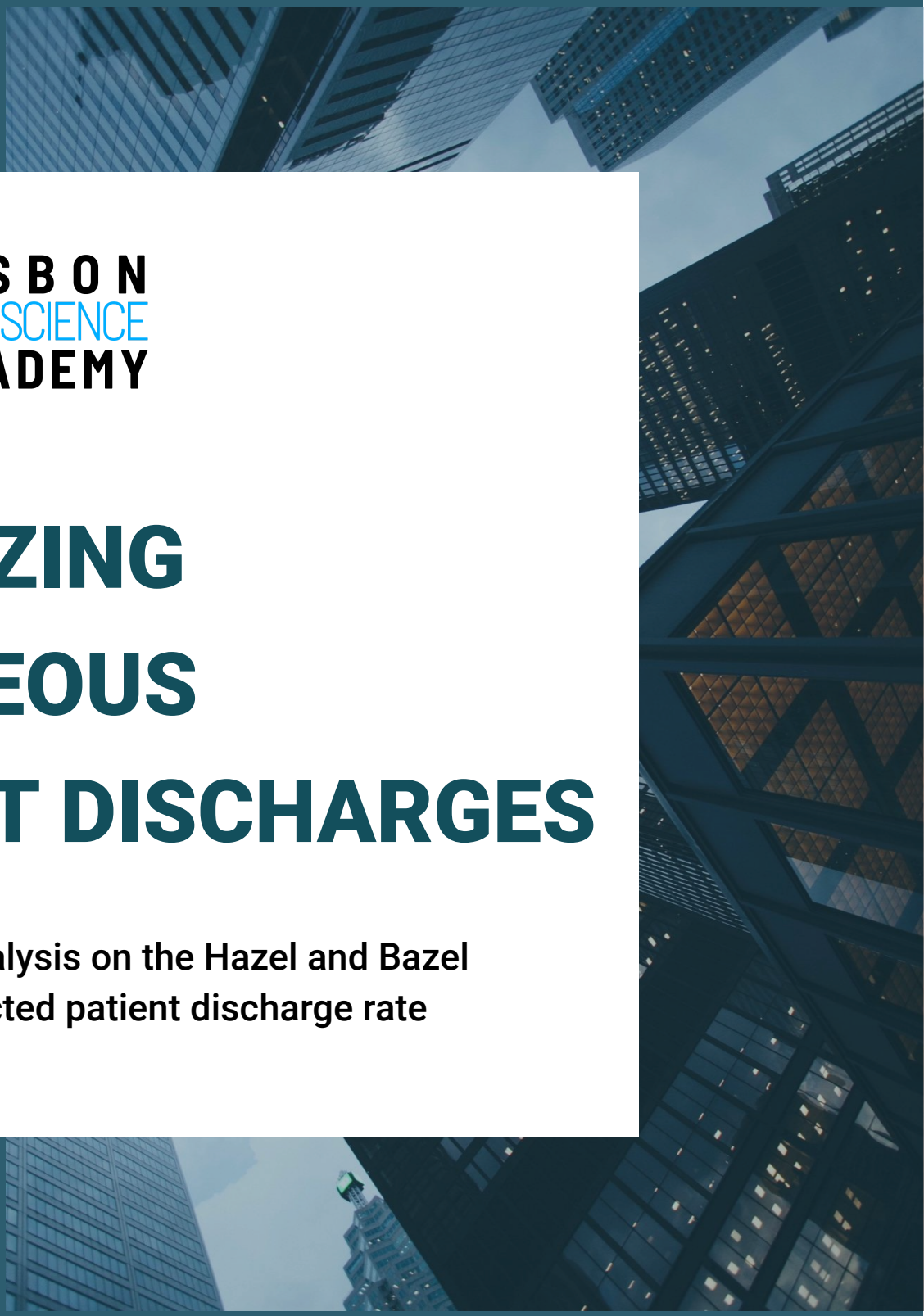


LISBON DATA SCIENCE ACADEMY

MINIMIZING ERRONEOUS PATIENT DISCHARGES

A data science analysis on the Hazel and Bazel Hospital's unexpected patient discharge rate

Prepared for:
Hazel and Bazel Hospital

Prepared by:
Rafael Gil
Data Scientist
Awkward Problem Solutions™

6 of March 2022

Contents

1 Client requirements

1.1 Summary

The Hazel and Bazel hospital, located in Los Angeles, California, is currently under investigation by the Fair Medical Practices Bureau due to suspicions of malpractice. The issue is related to the wrongful medical discharge of patients that return to the hospital within 30 days following the medical dismissal.

The hospital hired Awkward Problem Solutions™ to provide insights based on the data analysis of previous medical encounters. The main goal is to identify those patients who are likely to return to the hospital within 30 days following a medical discharge.

However, the analysis should also include an investigation into the reasons behind the unexpected patient discharge rate. In particular, if it is a demographic-specific issue or specific to some medical specialties or admission sources. Furthermore, the data analysis must also search for any evidence of discrimination based on gender, ethnicity, age, or insurance status regarding patients discharged from the Hazel and Bazel hospital.

Finally, Awkward Problem Solutions™ should also provide a REST API that integrates with the hospital's internal system. The API must validate patient discharges based on the medical encounter. In other words, the API should output whether to discharge a patient or not, based on the patient medical encounter information.

1.2 Requirements clarifications

The identified requirements can be divided into two different topics. First, there is the discrimination analysis on sensitive classes. Then, the predictive model deployment and its specifications.

Identifying discrimination, given its nature, is a subjective topic, specially when considering the discrimination on a per medical specialty basis on such a wide range of sensitive classes. There is a high probability of having multiple sub-classes under-represented due to the nature of the medical specialty itself, not discrimination.

Thus, this study will consider discrimination as a substantial difference between a sensitive class occurrence, e.g. asian race, in the whole dataset and a readmitted-only subset. The present study considers there is evidence of discrimination if the readmission rate, for a single sensitive class, is higher than 25 % of the class' overall occurrence.

Regarding, the predictive model, the main goal is to minimize wrongful discharge rate, hence the model should maximize its recall score. However, clarifications revealed that "at least 50 % of the patients identified for readmission should actually be sick". This translates into a minimum of 50 % precision score. Naturally, this requirement will impact the model recall score, since there is trade-off relationship between precision and recall. Thus, the predictive model should maximize the recall score whilst keeping a precision score above 50 %.

Finally, in order to ensure no discrimination by the predictive model, it is expected for the readmission rate not to vary more than 10 % between sub-groups and less than 5 % between medical specialties. The readmission rate can be measured with the precision score of of each sub-group and medical specialties. Thus, the precision score should not to vary more than 10 % between sub-groups and less than 5 % between medical specialties.

2 Dataset analysis

2.1 General analysis

The provided dataset contains information regarding a medical encounters at the hospital. There are 81 412 rows available, each representing a single medical encounter with 34 different variables. These variables describe the multiple aspects of a medical encounter, including patient information, diagnosis, and treatments. A detailed list of the variables included in the dataset is available on table 2.

Out of the 34 variables, `readmitted` is our explicit target as it indicates whether or not the patient returned to the hospital within 30 days of discharge. However, the number of encounters that resulted in readmission is only 11 % of the total, making the data at hand highly imbalanced. The remaining variables, described on tables 2 and 3 on section 5.1, ought to be considered as potential feature candidates.

2.1.1 Missing values

The vast majority of the categorical features have at least one or a combination of the following categories: `?`, `null`, `not mapped`, `not available`, `unknown/invalid`, `none`. These values were considered as missing and therefore converted into null before proceeding with the analysis.

Figure 1 lists all the variables available with their non-null percentage and non-null absolute count.

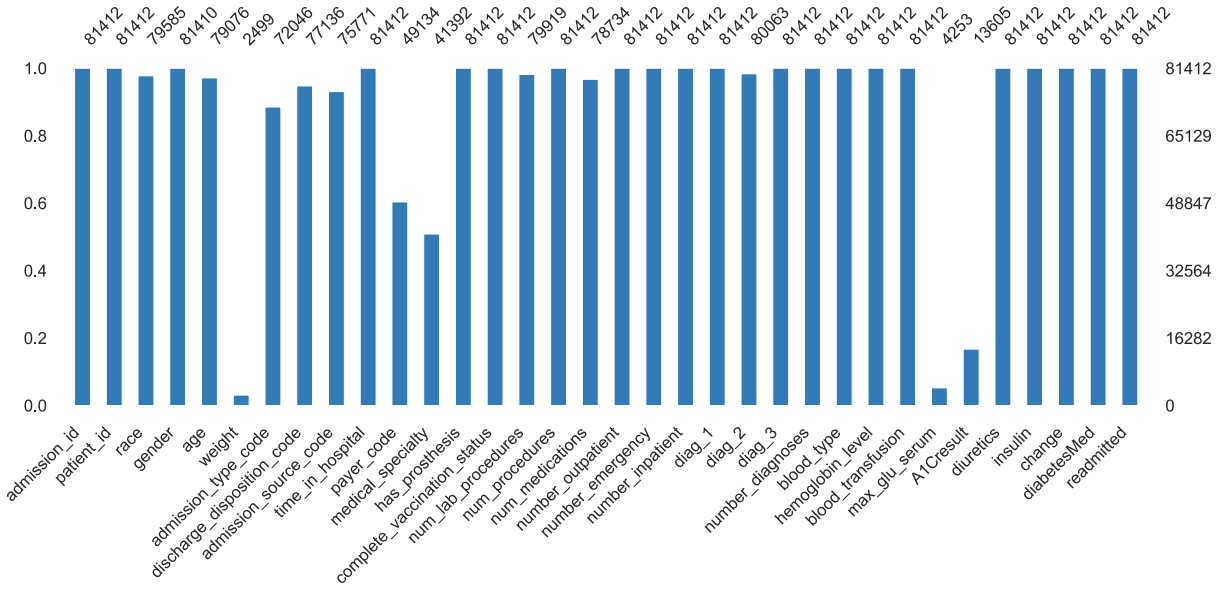


Figure 1: Medical encounter variables' availability

It is worth noting that `weight`, `max_glu_serum`, and `A1Cresult` are the worst offenders regarding nullability, with 96.9 %, 94.8 %, 83.3 % missing values respectively. Then, there is `medical_specialty`, `payer_code`, `admission_type_code`, `admission_source_code`, with 49.2 %, 39.6 %, 11.5 %, 6.9 % missing values respectively. There are a few more variables with a residual amount of missing values.

2.1.2 Number of unique values

The following variables have many different categories which will translate into a high dimension feature and may pose a burden on the model convergence:

`medical_specialty`**:** This variable identifies the medical specialty responsible for discharging the patient. The values available span through a range of 71 different medical specialties. However, there are cases where the main specialty is concatenated with a secondary specialty, resulting in a higher category count. For example: '*surgery cardiovascular*', '*surgery cardiovascular/thoracic*', '*surgery colon&rectal*', '*surgery general*', etc. Those were all merged into the same category '`Surgery`.' Furthermore, specialties with a count lower than 50 were merged into the same '`Others`' category, reducing the total category count to 24 different categories.

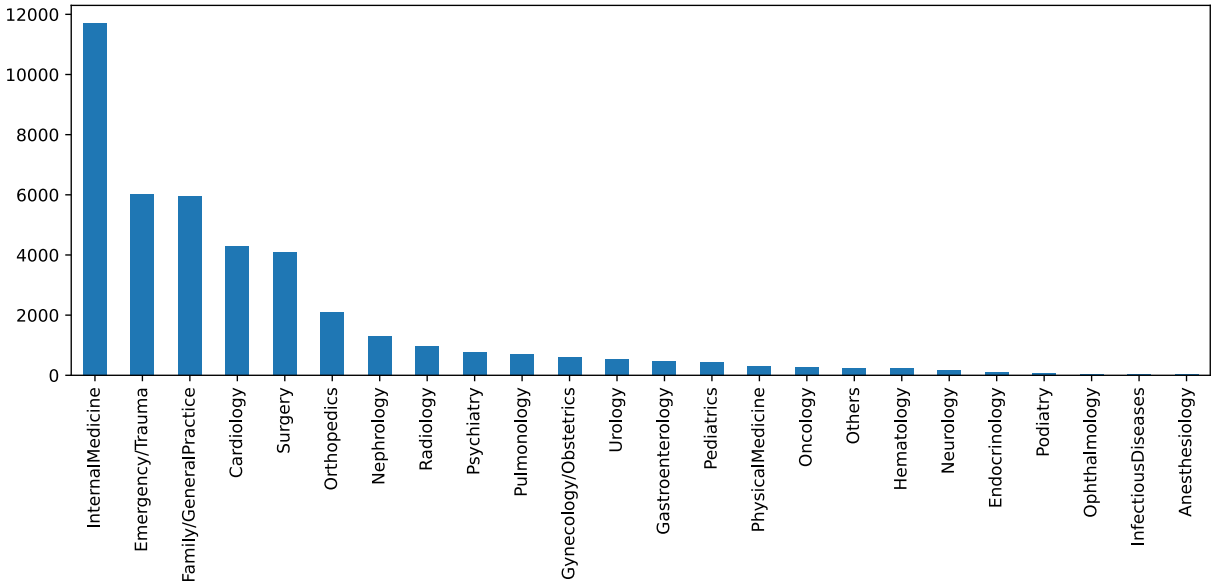


Figure 2: Number of medical encounters per medical specialty after processing

`diag_1, diag_2, diag_3`**:** These variables represent the ICD9 codes of the patient's primary, secondary and additional secondary diagnosis. The ICD9 is a list of codes for the International Statistical Classification of Diseases and Related Health Problems. The codes are organized into 18 different groups, and, to reduce the feature dimension, each code was converted into its group category. For instance, the codes ranging from 280 to 289 all belong to the group *diseases of the blood and blood-forming organs*, thus all these unique values were converted to their group name.

2.1.3 Correlation

Figure 3 shows the Spearman's correlation matrix of all the variables within the dataset. An initial look, reveals a strong positive correlation between `change`, `insulin`, and `diabetesMed`, with an highlight on `insulin` and `diabetesMed`. These are all boolean variables with the following meaning:

- `change` indicates if there was a change in diabetic medications;
- `insulin` indicates whether insulin was prescribed;

- `diabetesMed` indicates if there was any diabetic medication. prescribed.

These variables are all related to diabetes medication and empirically this relationship was expected, given the hospital's focus on patients with diabetes.

Furthermore, it is also worth noting a slight positive correlation between `num_medications` and `time_in_hospital`, as empirically expected.

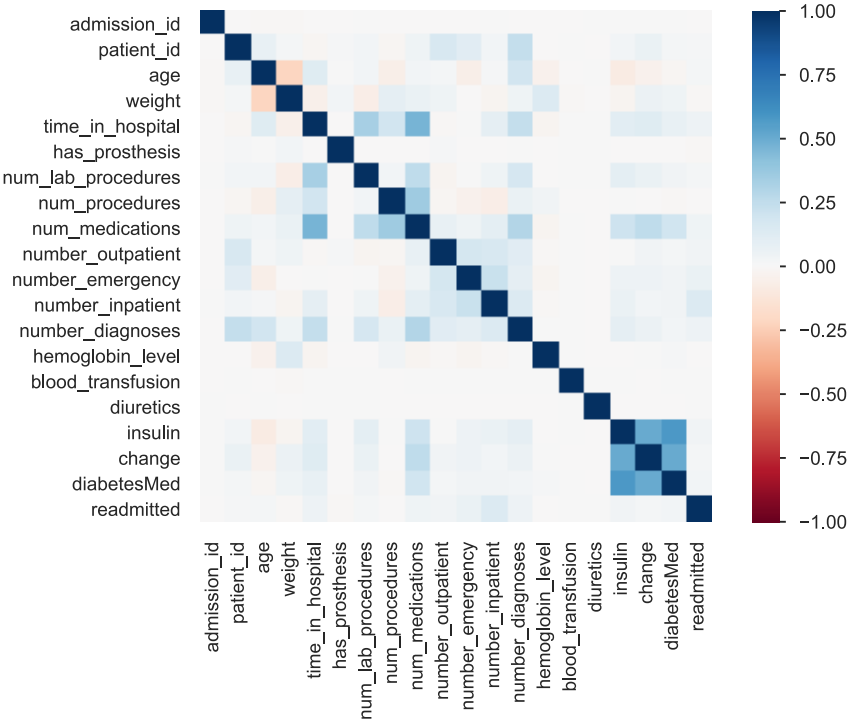


Figure 3: Spearman's correlation matrix

2.1.4 Data integrity

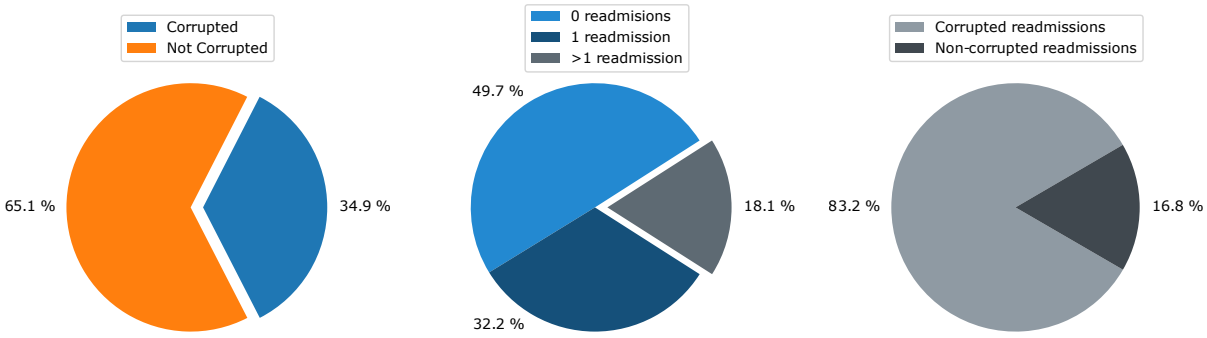
The provided dataset includes medical encounters of 60 069 unique patients. Out of those patients, 12 647, i.e. 21 % visited the Hospital more than once. This provides an opportunity to validate the dataset integrity since static patient information should be the same across all its medical encounters.

The variable `blood_type` was chosen to proceed with this analysis. `race`, and `gender` could also be argued as static patient information, but there is a degree of subjectivity to them and there is no chronological information available. Hence, the definition of a corrupted medical encounter: A medical encounter is corrupted if there is a second one where the same patient has a different `blood_type`.

Figure 4 summarizes the analysis to the whole dataset based on the corrupted medical encounter definition above. Figure 4a shows that 34.9 % of the data set is corrupted, i.e. each of these medical encounters have at least another one where `blood_type` does not match.

Figure 4b drills down into the corrupted entries and focus on the target variable: `readmitted`. It shows that 49.7 % of those corrupted medical entries belong to patients that were never readmitted into the hospital. 32.3 % belongs to patients that were readmitted once. The remaining, 18.1 % are entries of patients that were readmitted into the hospital at least twice.

Finally, figure 4c focus on the 18.1 % of corrupted medical entries that belong to patients readmitted more than once. Assuming that a medical encounter is handled with extra care



(a) Corrupted medical encounters across the whole dataset

(b) Target/Readmission distribution across the corrupted data

(c) Corrupted distribution on a readmitted-only subset from 4b

Figure 4: Data integrity analysis of based on patients who visited the hospital at least twice

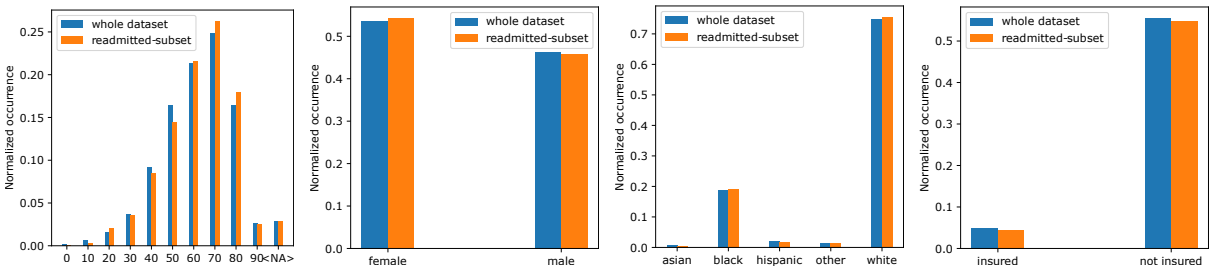
when a readmission occurs, and knowing that these patients were readmitted more than once, this chart checks if the all their readmitted medical encounters have a matching `blood_type`, i.e. if the re-admissions are corrupted when only compared to re-admissions. It turns out that only 16.8 % of those medical encounters have the same `blood_type` observation for the same patient. Of course, these 16.8 % medical encounters still have at least one non-readmitted observation with a different `blood_type` for the same patient, which is why they were considered corrupted in the first place on figure 4a.

2.2 Business questions analysis

There were concerns regarding patient discharge discrimination based on age, gender, race, and insurance status on the Hazel and Bazel hospital. To proceed with such analysis, patient discharge discrimination is hereby defined as substantial difference between a sensitive class occurrence in the whole dataset and a readmitted-only subset. In other words, there is evidence of discrimination if the occurrence of a sensitive class, e.g. asian race, is substantially higher on re-admissions when compared to the whole hospital's data.

Figure 5 details a discrimination analysis for the variables age, gender, race, and insurance status. In blue its plotted the occurrence of each sensitive class on the whole dataset, whilst in orange its plotted the occurrence of the same sensitive class across the readmission-only subset.

Empirically, from figure 5, there is no evidence of discrimination. There is only a slight issue of over-discharging the 70 years old age group and under-discharging the 50 years old age group, even though this is not an alarming discrepancy. Thus, the Hazel and Bazel hospital as whole does not discriminate on any of the studies variables when discharging patients.



(a) Age

(b) Gender

(c) Race

(d) Insurance status

Figure 5: Sensitive classes distribution across the whole dataset and a readmitted-only subset

However, the same analysis can be extended to each medical-specialty, thus revealing discrimination issues on a per-specialty basis. This allows finding out if there are discrimination issues on a single medical specialty that may have been hidden by the remaining medical specialties.

Analogously to the previous analysis, the discharge rate of the sensitive classes within a medical specialty was compared to the medical specialty's overall discharge rate. This study however, yielded different results. First, it was established a threshold of 25 % between the sensitive class occurrence and its discharge rate within a medical specialty. Assuming this threshold, there are many discrimination issues on different medical specialties. However, it is worth noting that the business requirements mention a lower 10 % threshold. Naturally, this results in higher discrimination occurrences, which shifts the focus from the worst offenders on this data analysis. Nonetheless, the processed data available here is normalized, making it agnostic to the threshold itself. Thus, this decision only impacts the highlighted medical specialties.

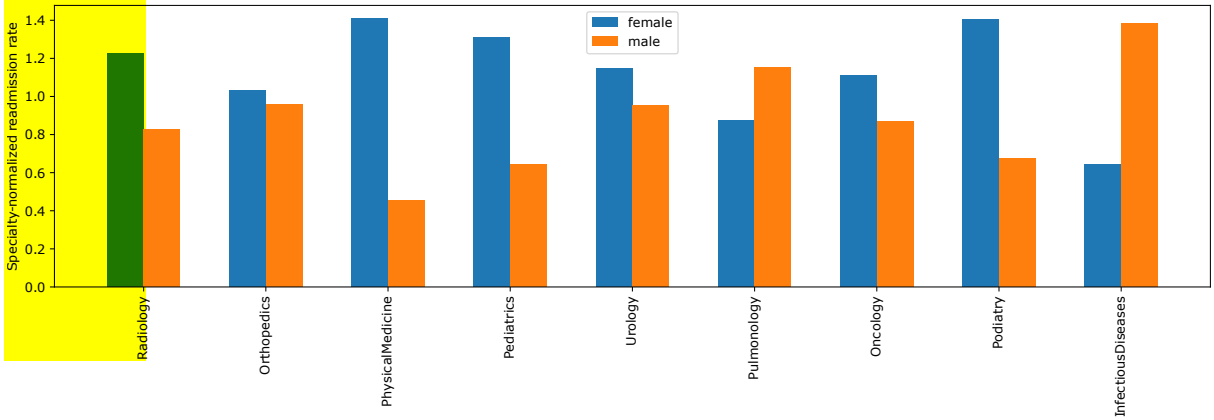


Figure 6: Gender discrimination analysis per medical specialty

Figure 6 summarizes the results of the per-specialty discrimination on the gender sensitive variable. The graph is built based on the data from table 4, available on section 5.1, along with the other variable's discrimination analysis. It shows, for instance, that the *Physical Medicine* medical specialty has readmission rate of female patients almost 40 % above the specialty's readmission rate. Meaning *Physical Medicine* is discriminating against women by early discharging female patients. A similar graphic is available on figures 10, 11, and 12, on section 5.2, regarding the remaining sensitive classes.

Given the analysis, and the discrimination definition above, every medical specialty, with the exception of *Internal Medicine* and *Family/General Practice*, is discriminating against at least one sensitive class.

2.3 Conclusions and Recommendations

Discrimination data seems alarming, but it does not mean there is structural discrimination on every specialization. For instance, *Pediatrics* shows a strong discrimination for age groups ranging from 60 years old to 80 years old, but this actually makes sense given the specialty's focus on lower age groups. Furthermore, medical specialties with lower registered medical encounters, such as *Endocrinology* — check its medical encounter count on figure 2 — have a higher probability of triggering precisely due to the low amount of medical encounters. This is enhanced by the fact that *Internal Medicine* and *Family/General Practice*, the non-discriminating medical specialties, are among the specialties with the most amount of medical encounters.

Also, the readmission discrimination analysis could be replicated for other variables such as `time_in_hospital`, `num_lab_procedures`, `number_diagnoses`, and `change`. This would allow identifying if the hospital is also discriminating on the patient's procedures and consequently strengthen the present analysis.

Finally, it is advised that the hospital focus on the specialties with the most medical encounters, that are indeed discriminating. For instance, the hospital is advised to verify the reasons behind *Emergency/Trauma* discrimination against young adults, the reasons behind the *Cardiology* discrimination against 10 year olds, or the reasons behind the *Orthopedics* discrimination against old patients and hispanics. Refer to table 4 for further details on the discrimination analysis.

3 Modelling

3.1 Model expected outcomes overview

The final classifier is able to identify a medical encounter that may result in a patient readmission within 30 days. According to the tests, it has an accuracy of 68 % and a recall of 79 %. 68 % accuracy means the model correctly identifies slightly more than 2 out of 3 medical encounters as a readmission or no-readmission, hence. On the other hand, 79 % recall means that out of the medical encounters that are readmitted, the model correctly identifies 79 % as readmissions. Recall is the measure of how good the model is at minimizing the erroneous patient discharge rate.

Recall, as mentioned, is the ability of the model to find all the medical encounters that will be readmitted, whilst precision is its ability not to mark as a potential readmission the medical encounters that won't be readmitted. Naturally, both these metrics have a trade-off relationship that is illustrated by figure 13.

The provided dataset, as discussed in section 2.1, is highly imbalanced meaning the amount of medical encounters that led to a readmission is a small fraction of the total dataset. The predictive models are highly sensitive to this issue and may not generalize as expected, resulting in real accuracy lower than the 68 % mentioned above. This accuracy has into account the requirement that at least 50 % of the patients identified for readmission are actually sick, whilst minimizing the number of erroneous medical discharges, i.e. maximizing recall whilst keeping precision above 50 %.

Furthermore, there are issues regarding the discrimination requirements and the reason is twofold. On one hand, there are few, i.e less than 100, medical encounters for each sensitive class, e.g asian race, hispanic race, etc. One the other hand, there is a wide range of sensitive classes, meaning there are a lot of different races and age groups. On top of that, there is also a wide range of medical specialties, further distributing the already low amount of classes throughout the medical specialties. Nonetheless, in an effort to reduce discrimination, the model was trained without access to these variables at the cost of a slightly lower performance.

As discussed in detail on section 3.3, the model is most sensitive to the `number_inpatient` variable, i.e. the number of inpatient visits of the patient in the year preceding the encounter.

3.2 Model specifications

In order to overcome the imbalanced dataset issue, it was under-sampled, since there there are over 80 000 records available. First, the corrupted non-readmitted medical encounters, identified in figure 4a of the dataset general analysis from section 2.1.4, were removed. Afterwards, a set of non-readmitted medical encounters was removed randomly until the target variable was finally balanced.

A different approach could be taken to reduce the amount of randomly removed samples. For instance dropping some specific `discharge_disposition_code` that may not represent a discharge from a medical facility. Nonetheless, this approach required further confirmation by the client.

Having the under-sampled dataset, the data was further divided into train and test data using a 80 %/20 % split, respectively. but keeping in mind that the target class should be evenly distributed among the two groups. Also, the target variable was extracted from each of the groups.

After this, the data is ready for the pipeline. The main focus was building a robust predictive model that could handle all kinds of input data, so a multi-step pre-processing was applied before the predictive model itself:

1. Pre-processing transformer - Convert input into valid data

   This step is responsible for converting raw data into the model's expected values and categories. The variables' types are checked and converted to an expected value if possible or `NaN` otherwise. Also, the categories are grouped and parsed at this staged to avoid duplicated and similar classes. There are also variables that are converted from categorical into numerical at this stage, such as `age`, `weight`, `hemoglobin_level`, `max_glu_serum`, and `A1Cresult`.

2. Column-dropper transformer

   This step is used as a handy and centralized location for column dropping. It was specially useful during the model's testing. Here the following columns are removed:

   - `admission_id` and `patient_id`, as mentioned in section 2.1, these are identifiers and have no predictive value;
   - `weight` and `max_glu_serum`, since both have over 90 % of missing values;
   - `insulin`, because it is highly correlated with `change` and `diabetesMed`
   - `age`, `gender`, `race`, ad `medical_specialty`, these are the sensitive variables, and they were removed and added interchangeably in order to evaluate their impact on the model's final discrimination.

3. Value imputer

   This step is useful to impute data that the first step could not. Categories that had missing values are filled with an extra `unknown` category. The remaining categorical and boolean variables with missing data are filled with the most common category. Numeric variables missing data, on the other hand, are imputed with their variable's mean.

4. One-hot Encoder / Standard Scaler

   On one hand, the one-hot encoder applies only to the categorical values and its goal is to convert those variables into numerical that the model can consume. On the other hand, the Standard Scaler is used to scale the numerical features exclusively.

5. Classifier

   Finally, the model is executed at the last stage of the pipeline. The chosen model, as discussed in section 3.4, is the *Gradient Boosting Classifier* with a maximum depth of 3, a total of 25 estimators, and a learning rate of 10 %. These hyper-parameters were chosen after running a grid search that provided the best F1 score.

After training the pipeline, it should calculate the probability of the two studied classes: `True` meaning the patient is expected to be readmitted, `False` meaning the patient is not expected to be readmitted. The last final step is to use this information to find the best deciding threshold that maximizes recall whilst keeping precision above 50 %, based on the test set predicted probabilities. The threshold that best fits the current model is 38 %. Figure 13 graphs the evolution of recall, precision, and accuracy when using different thresholds.

3.3 Analysis of expected outcomes based on training set

The evaluated models are all prune to over fitting, meaning they easily memorize the training data and fail to generalize the problem. To handle this issue the number of estimators and

the model's max depth were reduced, thus forcing the model to focus on the most important variables.

Figure 7 plots the feature importance of the final model. As discussed the most important feature is the `number_inpatient`, with over 30 % importance. Followed by `number_outpatient`, `number_emergency`, and `number_diagnoses`.

The model has an recall of 79 % whilst keeping a precision above 65 %. In other words, the model correctly identifies 79 % of the medical encounters as readmissions out of those that are actually readmitted, whilst making sure that at least 65 % of the readmitted are actually sick.

This is achieved with a threshold of 38 %, meaning the model has to output a probability of readmission higher than 38 % for the model to consider it as a readmission. This threshold is the one that maximizes recall keeping precision above 65 %.

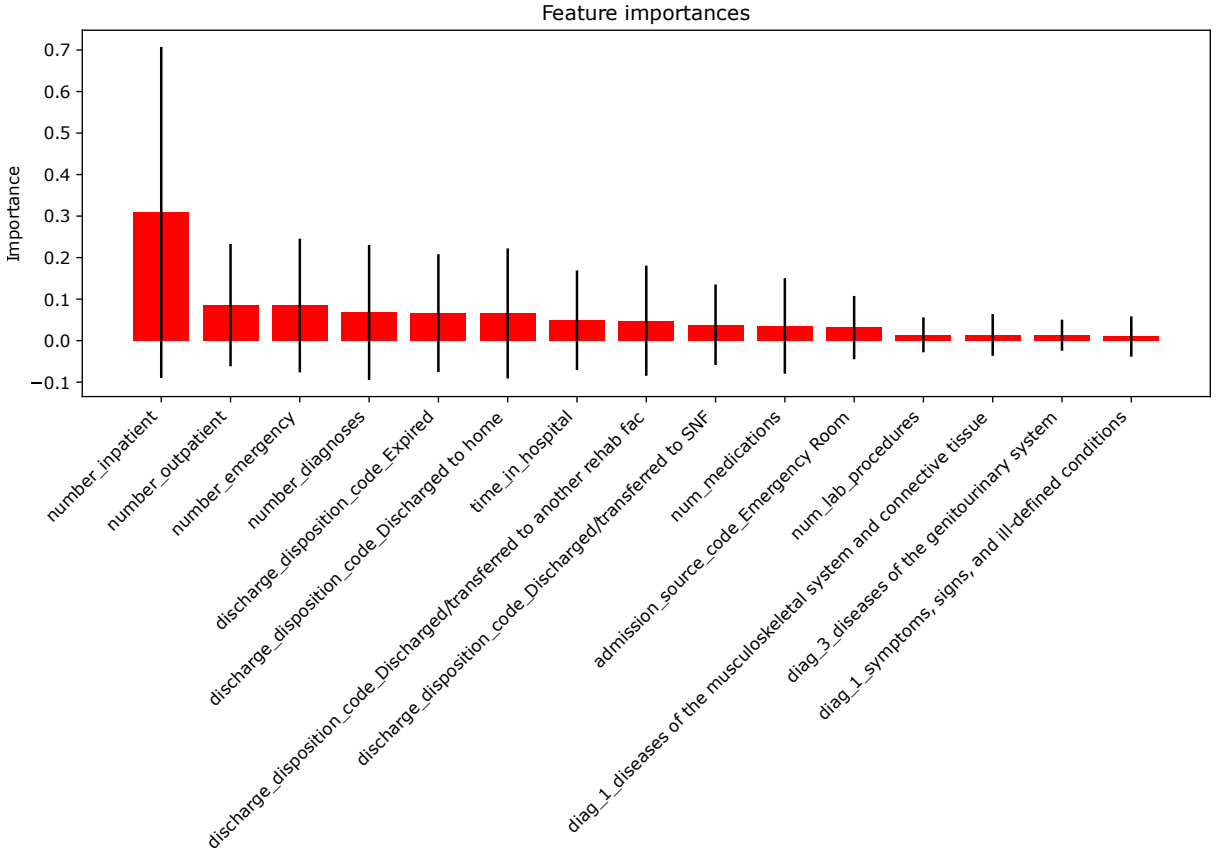


Figure 7: Final model's feature importance

Regarding discrimination, the business requirements state that the model should keep the readmission rate variance below 10 % between sub-groups and below 5 % between medical specialties. As mentioned, the readmission rate can be measured through the precision score of each sub-group and medical specialties. Figure 14 plots the model's precision score for the different sensitive classes, whilst figure 15 plots the model's precision scores for the different medical specialties.

The results do not meet all criteria mainly due to the low amount of observations on the test set, and consequently on each of the sensitive classes, due to the undersampling. Two sub-groups meet the criteria: gender, and insurance status with precision scores ranging a maximum of 4.86 % and 4.73 %, respectively. However, the criteria is not met for age, and race with with precision scores ranging a maximum of 36.66 % and 19.05 %, respectively. Note that age and

race are sensitive variables with a higher number of categories, and consequently each category has a lower amount of observations. A low amount of observations produces spikes on the precision score that render this discrimination requirement difficult to overcome. The medical specialty requirement also did not meet the requirement with a precision variance of 47.29 %. There are even more medical specialties than age and race categories, thus the same rationale applies.

The sensitive classes were removed from the training data in an effort to reduce explicit discrimination. Thus, the model has no access to those variables when predicting a result. Unfortunately, this does not mean there won't be discrimination. Nonetheless, higher volumes of data may reveal the discrimination is lower than the results now reveal.

3.4 Alternatives considered

Identifying whether a patient is expected to be readmitted within a 30 day period is a classification problem. Thus, there are several predictive models available that better fit the current conundrum, such as the *Random Forest Classifier*, the *Logistic Regression*, the *Stochastic Gradient Descent*, and the *Gradient Boosting Classifier*.

Nonetheless, there are client requirements that impose a minimum 50 % recall as detailed in section 1.2. Naturally, the goal of a predictive model is to maximize its precision, however recall and precision have a trade-off relationship. As a result, the selected model should provide the highest precision that still allows the required minimum recall.

Given the current scenario, the F1 score is useful roughly to assess the best model, since it measures the the harmonic mean of precision and recall through the equation 1.

$$F1_{score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

Thus, before committing into a model each of the models listed above was evaluated through the F1 score. Results are displayed on figure 8 and its source is available on table 5 of section 5.3.

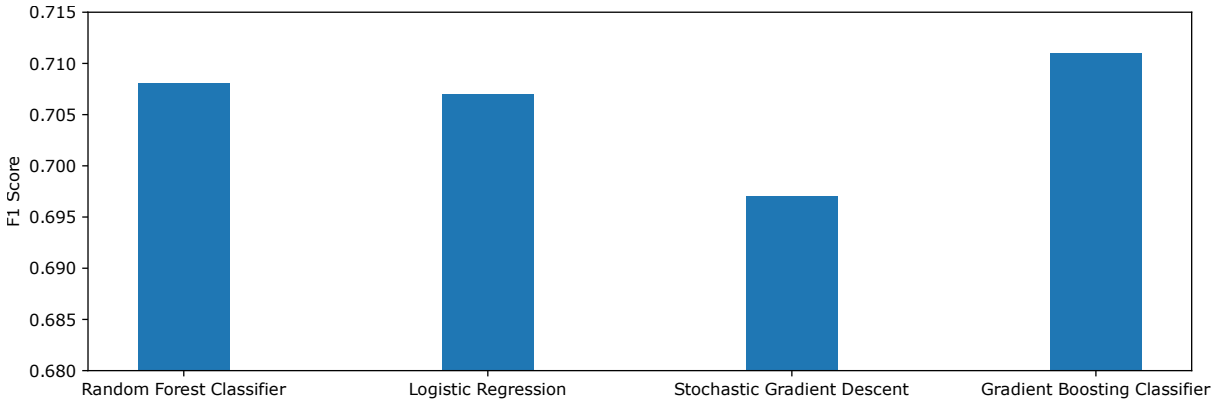


Figure 8: Model's best F1 score on the test data set

The results are close across all the models, but the *Gradient Boosting Classifier* had a slight edge over the remaining. For that reason, the *Gradient Boosting Classifier* was chosen to proceed the analysis.

3.5 Known issues and risks

The current approach is highly vulnerable to over-fitting, which is the main risk of the proposed model. This means the model may not be able to provide an accurate result for data it has never seen, specially when a medical encounter has details that were not available during training.

Also, the high dependency on a single variable may cause the model fail for observations where that variable is missing.

Furthermore, *Gradient Boosting Classifier* is sensitive to outliers given the logic it uses to build its decision trees. This means that outliers on our current data set can have higher impact on the model performance. This problem is exacerbated by the random under-sampling used to balance the data, since outliers on the training set have a higher preponderance on the model.

The proposed model also fails to avoid discrimination, even though there are constraints in place to mitigate it. A higher volume of data may be useful to improve on this issue on a future iteration.

4 Model Deployment

4.1 Deployment specifications

The predictive model was built taking into consideration its deployment and tried to be as generic as possible regarding the input data. This was achieved using the `sklearn` framework as it provides useful tools and boilerplate for this endeavour. The model predictor is built within a `pipeline` that also contains `transformers`. More specifically, all variables from the provided dataset have its own `transformer` that either drops the variable or converts it into a valid input. Furthermore, `sklearn`'s `imputers` were also used to impute default values on variables that may be missing. The combination of `transformers` and `imputers` allowed having a `pipeline` ready to handle a wide range of inputs, thus contributing to the overall application robustness.

The model deployment was also built to be as robust as possible. Even though the schema API has all the parameters marked as mandatory, the model deployment took into consideration edge cases that could crash the application. Thus, all the request fields can be missing and an output is still provided.

In order to keep track of the model performance and enable future deployments with up to date data, the application is integrated with a `postgres` database. This database contains a single table called `MedicalEncounter` that matches the raw data of the provided dataset according to table 2. Each variable on that table has its own column with a matching data type: either `Text`, `Integer`, or `Float`. Furthermore, 3 extra columns were added:

- `proba`: a `Float` field to hold the model probability output for the current observation;
- `prediction`: a `Boolean` field to hold the actual model prediction based on the probability;
- `true_label`: a `Boolean` field to store the actual expected outcome;

The database connection and queries are handle with `pewee`, a ORM (Object–relational mapping) library responsible for interpreting `python` instructions and data into the database equivalent, and vice-versa. It is used to connect with the database and build a representation model of the stored data.

The request data integrity is also assured using `pewee`. Hence, why all expected fields are marked as either `text`, `float` or `boolean` with the advantage of `pewee` itself doing the type conversions. For instance, `pewee` recognizes the string *"11"* as the number 11 and still stores it as a `float`. When `pewee` fails to do the type conversion, the database will throw an error if the data type is invalid. Such exception is handled and modified into an error response message readable by the end user.

The web application was built using the python's library `flask` and provides two different endpoints as described in table 1. When an error occurs, for instance when there is data type mismatch as described above, both endpoints return an error response whose schema is detailed on algorithm 5.

Naturally, errors may occur on both endpoints, and consequently there is an error response schema, detailed on algorithm 5, used to handle those cases. The application has 4 different error types:

- `INVALID_REQUEST`: Triggered when a request does not match the expected format and it is not possible to infer the correct format.
- `DUPLICATED_ADMISSION_ID`: Triggered for duplicated `/predict` requests. This error message is not exposed outside the application of the application, since the app will recover from it and return the previous known prediction;

Table 1: REST API endpoints description

Method	Path	Description
`POST`	`/predict`	Receives information regarding a single medical encounter, and outputs whether the patient should or should not be discharged. The request schema is detailed on algorithm 1 and the response schema on algorithm 2. All unique prediction requests are stored in a database for future analysis.
`POST`	`/update`	Receives one medical encounter identifier along with the real readmission information. The database should be updated accordingly for future reference. The request schema is detailed on algorithm 3 and the response schema on algorithm 4.

- `DOES_NOT_EXIST`: Triggered when trying to update a medical encounter that does not exist in the database;
- `UNKNONW_EXCEPTION`: General purpose error response. This error should never be triggered, since it is used for unpredicted runtime errors;

Finally, the whole application is deployed and published through `Heroku`. `Heroku` also provides the database resources used to collect the data throughout the application lifespan.

4.2 Known issues and risks

Event though robustness was one of the key factors in the application deployment, there are a few details worth noting. First of all, the application is not fault tolerant regarding the request's field names. This means, that a typo on the request is interpreted as a missing field. Missing fields, however, will not impact the application's health, but may skew the model correctness.

Furthermore, the application discards duplicated medical encounters and keeps only the first occurrence. This may be an issue if the first request of a given encounter has missing information that is available on a second request.

Also, the application is 100 % dependent on the `Heroku` service provider availability. If for any reason, `Heroku`'s services are down or compromised, the REST API follows suit. Furthemore, `Heroku` imposes limitations on the application usage:

- Maximum 10 000 unique records on the database;
- Maximum of 4500 requests per hour;
- The logging availability is limited to the last 1500 lines.

Regarding security, the data provided to the REST API is stored in clear text on the data base, including sensitive patient data. This means, that an attacker with elevated access to our system could download these information. Once more, the app relies entirely on `Heroku` security to handle access authorization. However, this also means that a breach on their systems could also comprise the data stored by the REST API.

Finally, `Heroku` sets its applications to sleep if they receive no traffic within 1 h. When that happens, it is expected for the first requests to suffer a small delay. Nonetheless, all requests should perform as expected after the application has wakened.

Algorithm 1 `/predict` request schema

```
{
  "admission_id": 0,
  "patient_id": 0,
  "race": "string",
  "gender": "string",
  "age": "string",
  "weight": "string",
  "admission_type_code": 0,
  "discharge_disposition_code": 0,
  "admission_source_code": 0,
  "time_in_hospital": 0,
  "payer_code": "string",
  "medical_specialty": "string",
  "has_prosthesis": true,
  "complete_vaccination_status": "
   string",
  "num_lab_procedures": 0,
  "num_procedures": 0,
  "num_medications": 0,
  "number_outpatient": 0,
  "number_emergency": 0,
  "number_inpatient": 0,
  "diag_1": "string",
  "diag_2": "string",
  "diag_3": "string",
  "number_diagnoses": 0,
  "blood_type": "string",
  "hemoglobin_level": 0,
  "blood_transfusion": true,
  "max_glu_serum": "string",
  "A1Cresult": "string",
  "diuretics": "string",
  "insulin": "string",
  "change": "string",
  "diabetesMed": "string",
  "readmitted": "string"
}
```

Algorithm 2 `/predict` response schema

```
{
  "readmitted": "string"
}
```

Algorithm 3 `/update` request schema

```
{
  "admission_id": 0,
  "readmitted": "string"
}
```

Algorithm 4 `/update` response schema

```
{
  "admission_id": 0,
  "actual_readmitted": "string",
  "predicted_readmitted": "string"
}
```

Algorithm 5 Error response schema

```
{
  "detail": [
    {
      "loc": [
        "string"
      ],
      "msg": "string",
      "type": "string"
    }
  ]
}
```

Figure 9: Endpoints' request, response and error schema

5 Annexes

5.1 Dataset technical analysis

Table 2: Dataset variables' types before and after pre-processing

Variable	Original Type	Pre-processed Type
admission_id	int64	int64
patient_id	int64	int64
race	object	category
gender	object	category
age	object	Int64
weight	object	Int64
admission_type_code	float64	category
discharge_disposition_code	float64	category
admission_source_code	int64	category
time_in_hospital	int64	int64
payer_code	object	category
medical_specialty	object	category
has_prosthesis	bool	bool
complete_vaccination_status	object	category
num_lab_procedures	float64	Int64
num_procedures	int64	int64
num_medications	float64	Int64
number_outpatient	int64	int64
number_emergency	int64	int64
number_inpatient	int64	int64
diag_1	object	category
diag_2	object	category
diag_3	object	category
number_diagnoses	int64	int64
blood_type	object	category
hemoglobin_level	float64	float64
blood_transfusion	bool	bool
max_glu_serum	object	category
A1Cresult	object	category
diuretics	object	bool
insulin	object	bool
change	object	bool
diabetesMed	object	bool
readmitted	object	bool

Table 3: Numerical variables description after pre-processing

Variable	count	mean	std	min	25 %	50 %	75 %	max
age	79076	60.96	15.97	0	50	60	70	90
weight	2499	73.67	26.23	0	50	75	100	200
time_in_hospital	81412	4.40	2.98	1	2	4	6	14
num_lab_procedures	79919	43.07	19.63	1	32	44	57	132
num_procedures	81412	1.34	1.71	0	0	1	2	6
num_medications	78734	16.02	8.11	1	10	15	20	81
number_outpatient	81412	0.37	1.28	0	0	0	0	42
number_emergency	81412	0.20	0.88	0	0	0	0	64
number_inpatient	81412	0.64	1.27	0	0	0	1	21
number_diagnoses	81412	7.42	1.93	1	6	8	9	16
hemoglobin_level	81412	14.19	1.06	10	13	14	15	19

5.2 Business questions technical support

Table 4: Detailed per-specialty discrimination analysis. All values are normalized to the overall medical specialty readmission rate on the last column. Values above the 25 % threshold are marked in **bold**, values below the threshold are underlined. Dashes mean there were no occurrences for that class, whilst zeros mean there no re-admissions for that class.

Medical Specialty	Age											Gender		Race					Insured		Spec. Rate
	0	10	20	30	40	50	60	70	80	90	NA	F	M	asian	Black	Hisp.	Other	White	Yes	No	
Emergency/Trauma	0	0	**2.74**	**1.31**	1.05	0.67	0.92	0.90	1.13	1.24	**1.35**	1.05	0.94	0.57	0.86	0.86	0.81	1.04	0.85	1.04	0.11
InternalMedicine	-	0.60	1.07	1.20	0.87	0.83	1.02	1.11	1.03	0.79	1.03	1.02	0.98	0.82	1.01	1.21	0.84	1.00	0.89	1.00	0.11
Family/GeneralPractice	0	0	0.67	0.93	1.06	0.75	1.04	1.10	1.13	0.93	0.79	0.99	1.01	1.10	1.03	1.07	0.53	0.99	1.03	1.00	0.12
Radiology	-	0	**5.19**	0	**1.44**	0.99	0.83	0.91	**1.71**	0	0.37	1.23	0.83	0	1.18	**1.30**	**2.97**	0.96	1.21	0.99	0.10
Orthopedics	-	-	0	0.43	0.67	0.80	0.87	**1.31**	1.06	**1.61**	**1.31**	1.03	0.96	0	1.19	**2.11**	0	0.98	0.23	1.02	0.09
Cardiology	0	**4.20**	0	0.42	0.78	0.87	1.01	1.06	1.22	**1.40**	1.13	0.95	1.04	0	1.01	0.59	1.22	1.01	0.91	1.00	0.08
PhysicalMedicine	-	-	-	0	0.30	1.16	0.57	1.11	**1.42**	**2.23**	**1.49**	**1.41**	0.46	0	**1.65**	**3.35**	0.96	0.83	0	1.01	0.15
Psychiatry	-	0.99	0.72	1.07	0.86	0.92	0.96	**1.51**	**1.37**	1.11	0.42	0.87	1.21	0	0.78	0	1.11	1.12	0.47	1.01	0.11
Surgery	0	0	0.72	0.77	0.97	1.10	0.92	1.02	0.96	**1.34**	**1.49**	0.95	1.05	0.77	0.96	**1.26**	1.13	1.01	1.03	1.00	0.10
Hematology	-	-	0	**4.34**	**1.45**	0.75	**1.38**	0.84	0.89	0.62	0.54	0.88	1.11	**3.25**	1.08	-	**1.81**	0.88	**1.45**	0.98	0.23
Others	**15.00**	-	0	0	**1.30**	0.68	0.79	0.92	1.07	0	**4.29**	**1.30**	0.66	0	0.65	**1.67**	0	1.11	0	1.03	0.07
Gynecology/Obstetrics	-	0	**1.31**	0.74	**1.49**	**1.26**	0.91	0	0	0	**1.28**	1.00	-	0	**1.83**	0.66	0	0.70	1.02	1.00	0.05
Pediatrics	0	0.72	0	0	**1.60**	0.77	**6.21**	**3.05**	**3.19**	0	**2.03**	**1.31**	0.64	0	**1.27**	**2.79**	0	0.85	0	1.00	0.04
Nephrology	-	0	**2.82**	1.07	1.10	0.95	1.01	0.93	0.91	0.45	0.88	1.00	1.00	0	0.94	0.97	0	1.09	1.15	0.99	0.16
Urology	-	-	**3.27**	0.70	0.83	**1.57**	0.96	0.96	0.44	0	0.75	1.14	0.95	**3.67**	0.79	0	0.89	1.04	**1.81**	0.96	0.10
Anesthesiology	0	0.87	-	-	0	0	-	**6.50**	-	-	0	0.76	**1.44**	-	0	0	-	**1.62**	-	1.00	0.08
Gastroenterology	-	-	0	0.98	**1.88**	0.58	0.85	1.13	0.98	1.06	0	1.08	0.93	0	0.76	0	1.06	1.06	**1.69**	0.98	0.12
Pulmonology	-	0	0	**1.37**	0.45	0.84	0.89	1.13	**1.32**	**1.31**	**1.53**	0.87	1.15	**2.02**	0.73	0.56	**1.74**	1.06	1.01	1.00	0.10
Oncology	-	0	-	0	**1.26**	**1.28**	0.79	0.85	1.20	-	0.88	1.11	0.87	-	0.92	0	**2.64**	1.04	0	1.00	0.19
Podiatry	-	-	-	**2.59**	0	0.69	**1.43**	0	**1.73**	0	**2.08**	**1.40**	0.68	0	**2.59**	-	0	0.79	0	1.04	0.10
Neurology	0	-	-	0	0.74	0.92	**1.28**	**1.32**	1.07	0	0	1.16	0.88	0	0.83	0	0	1.05	0.90	1.01	0.06
Endocrinology	-	0	0	0	**2.38**	0	0	0	**7.41**	0	0	0.64	**1.39**	-	0	**33.33**	-	0.89	0	1.01	0.03
Ophthalmology	-	-	-	0	0	0	**1.42**	**1.89**	0	-	-	0.89	1.13	-	0	-	-	**1.31**	-	1.00	0.06
InfectiousDiseases	-	-	-	0	0.69	0	0	**3.22**	**2.42**	-	-	0.64	**1.38**	**4.83**	0	**4.83**	-	0.81	-	1.00	0.21

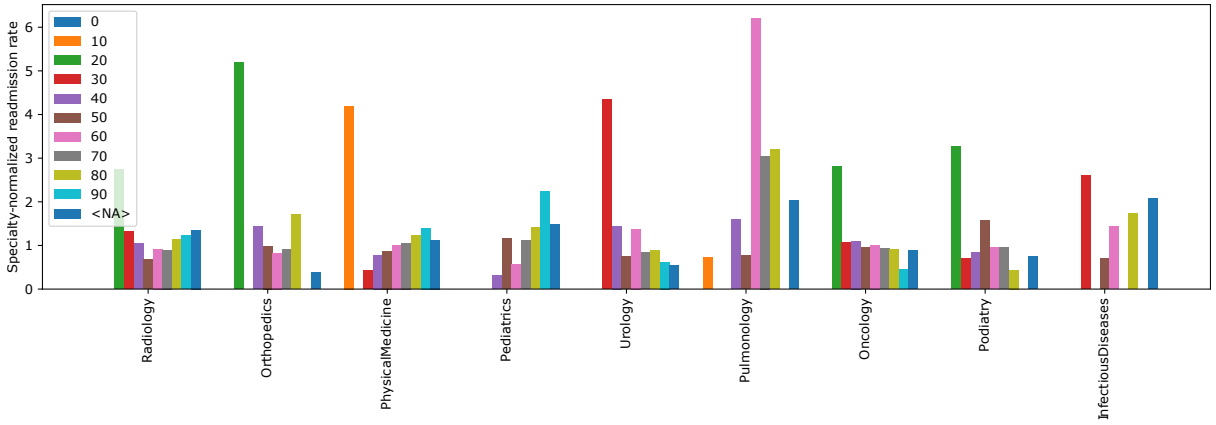


Figure 10: Age discrimination analysis per medical specialty

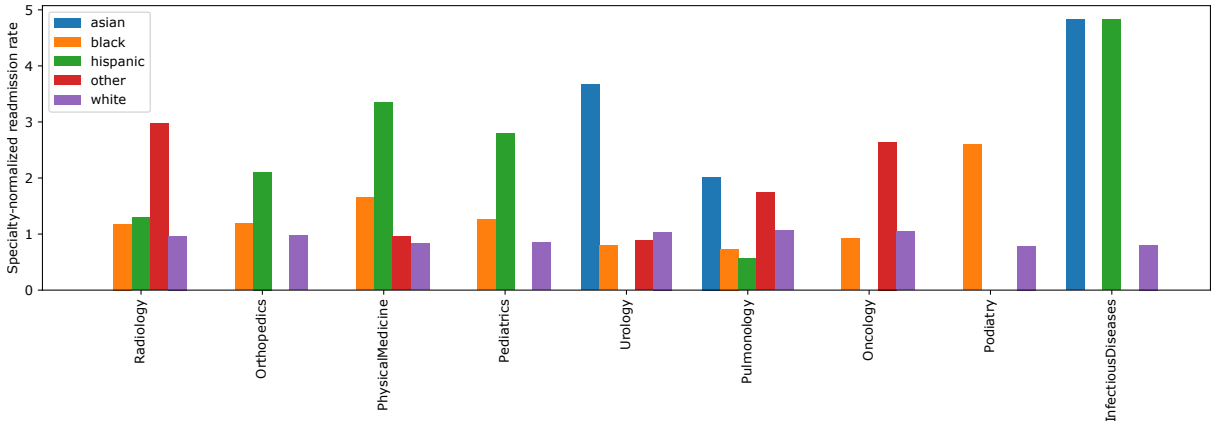


Figure 11: Race discrimination analysis per medical specialty

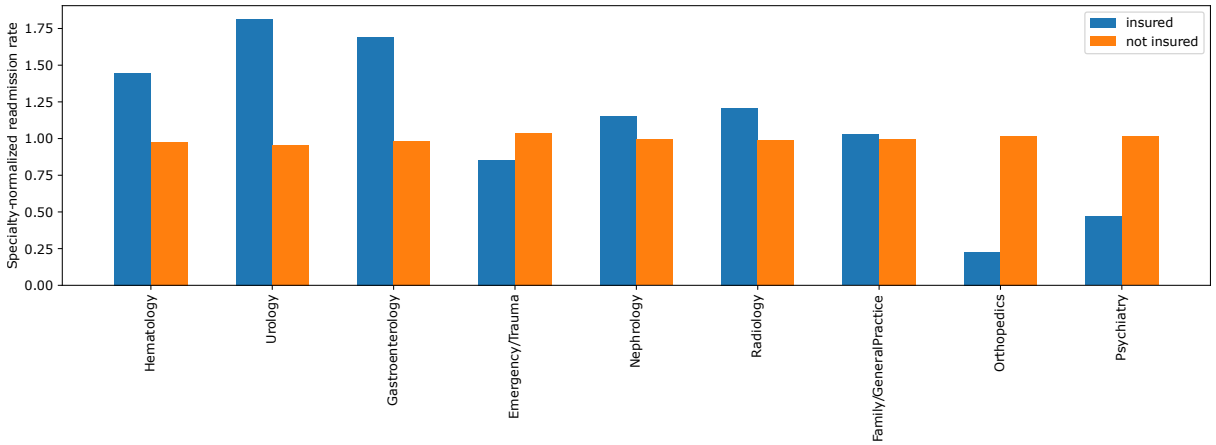


Figure 12: Insurance status discrimination analysis per medical specialty

5.3 Model technical analysis

Table 5: Predictive model's best F1 score on the test data set

Model	F1 Score
`Random Forest Classifier`	0.708
`Logistic Regression`	0.707
`Stochastic Gradient Descent`	0.697
`Gradient Boosting Classifier`	**0.711**

Table 6: Predictive model's threshold that maximizes precision but keeping recall higher than 50 % both on the training and the test data set, using the default hyper-parameters

	Model	Threshold	Precision	Recal	F1
Train	`Random Forest Classifier`	0.85	**1.000**	**0.520**	**0.685**
	`Logistic Regression`	0.64	0.835	0.502	0.627
	`Stochastic Gradient Descent`	0.64	0.821	0.509	0.629
	`Gradient Boosting Classifier`	0.65	0.832	0.507	0.630
Test	`Random Forest Classifier`	0.60	0.794	0.500	0.614
	`Logistic Regression`	0.61	0.797	0.500	0.614
	`Stochastic Gradient Descent`	0.61	0.791	**0.510**	**0.620**
	`Gradient Boosting Classifier`	0.62	**0.798**	0.501	0.616

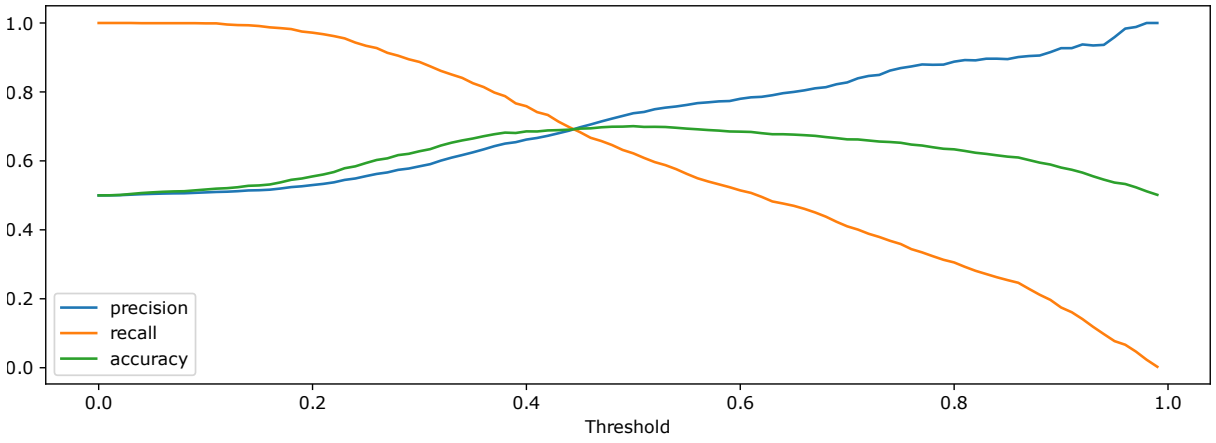
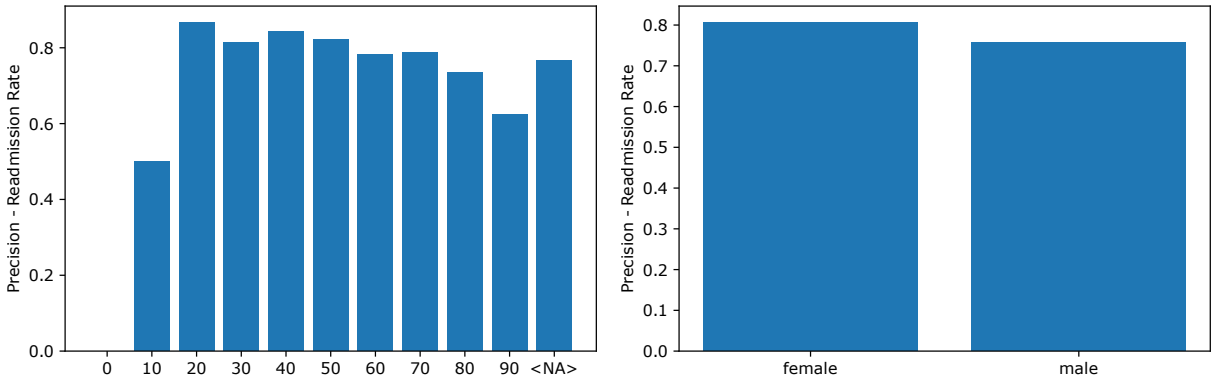
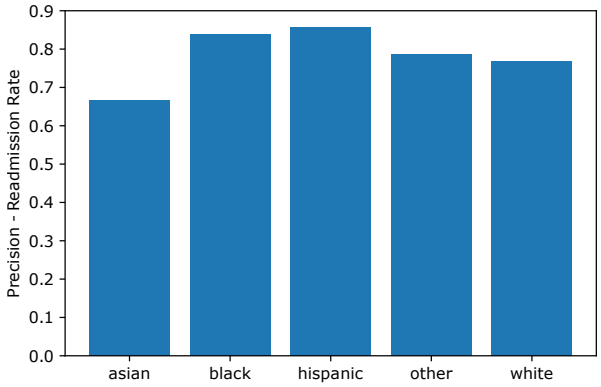


Figure 13: Precision, recall, and accuracy evolution as the threshold increases

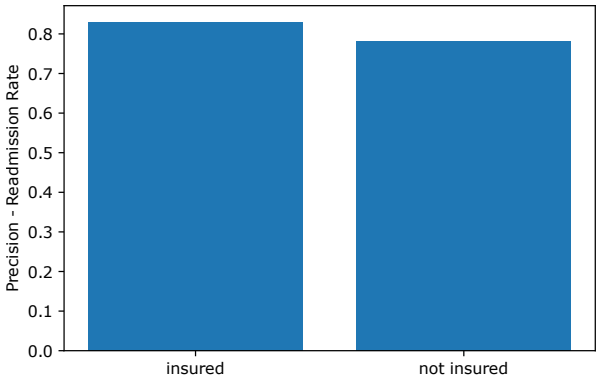


(a) Age

(b) Gender



(c) Race



(d) Insurance status

Figure 14: Model readmission rate for sensitive sub-groups

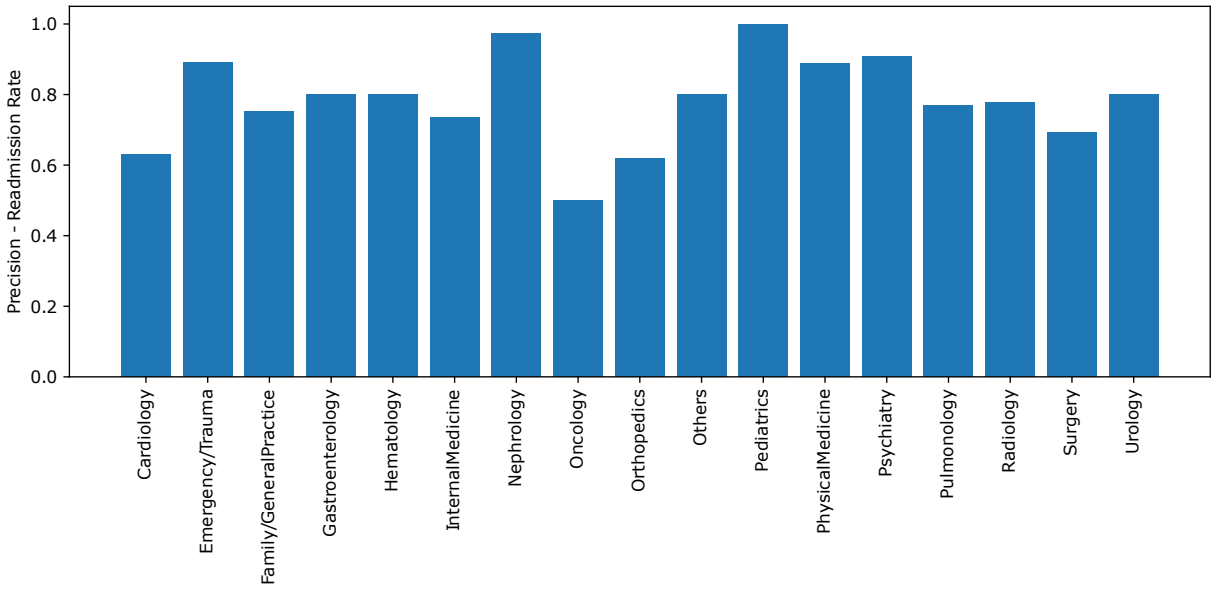


Figure 15: Model readmission rate for medical specialties

Table 7: *Gradient Boosting Classifier* detailed feature importance graphically displayed on figure 7

Feature Name	Importance
number_inpatient	30.87 %
number_outpatient	8.55 %
number_emergency	8.45 %
number_diagnoses	6.79 %
discharge_disposition_code_Expired	6.62 %
discharge_disposition_code_Discharged to home	6.55 %
time_in_hospital	4.90 %
discharge_disposition_code_Discharged/transferred to rehab fac	4.80 %
discharge_disposition_code_Discharged/transferred to SNF	3.84 %
num_medications	3.53 %
admission_source_code_Emergency Room	3.14 %
num_lab_procedures	1.39 %
diag_1_diseases of the musculoskeletal system	1.37 %
diag_3_diseases of the genitourinary system	1.30 %
diag_1_symptoms, signs, and ill-defined conditions	0.99 %
payer_code_MC	0.80 %
payer_code_BC	0.61 %
diag_2_diseases of the skin and subcutaneous tissue	0.57 %
admission_source_code_Clinic Referral	0.44 %
admission_source_code_Physician Referral	0.43 %
A1Cresult	0.41 %
discharge_disposition_code_Discharged/transferred to home	0.32 %
discharge_disposition_code_Discharged/transferred to another	0.27 %
num_procedures	0.26 %
hemoglobin_level	0.25 %
diag_1_external causes of injury	0.21 %
diag_3_endocrine, nutritional and metabolic diseases	0.16 %
diag_2_complications of pregnancy, childbirth	0.16 %
diag_2_infectious and parasitic diseases	0.14 %
discharge_disposition_code_Hospice / medical facility	0.12 %
admission_type_code_Urgent	0.12 %
admission_type_code_Emergency	0.11 %
admission_source_code_Transfer from a hospital	0.10 %