LISBON DATA SCIENCE ACADEMY

ANALYSIS AND PREDICTION OF HOSPITAL DISCHARGES

MACHINE-LEARNING-BASED PREDICTION OF POTENTIALLY WRONGFUL HOSPITAL DISCHARGES

Prepared for:
Hazel and Bazel Hospital

Prepared by:
Ana Milas
Data scientist
Awkward Problem Solutions

6 of February 2022

Table Of Contents

1. Client requirements

1.1 Summary

The client is Hazel and Bazel Hosipital, a hospital located in Los Angeles, California, with special focus on treating diabetes. The client is concerned about the rates of early readmissions of the patients (less than 30 days after discharge), and set out two main requirements.

First requirement was to perform the analysis of available data to assess whether there is any discrimination based on gender, ethnicity, age or insurance status of the patient. Furthermore, the client requested to analyze whether any of their specialties are having disproportionately high rates of premature discharges.

Second requirement was to develop a model that will predict if the patient is likely to be readmitted in the 30 days following the discharge. The client would use this model to prolong the stay of the patients that are likely to be readmitted. The model should be sensitive enough to detect most of the potential premature discharges. However, due to limited resources of the hospital, it should not have a high rate of identifying the healthy patients as potential readmissions. Furthermore, this model should not discriminate based on gender, ethnicity, age or the insurance status of the patient. Finally, the model is to be delivered in the form of REST API that could be easily integrated in the hospital's internal system for medical discharge approval.

1.2 Requirements clarifications

Following the initial email of the client, further clarification of the business requirements was requested in order to turn them into data science metrics.
The client requested that at least 50% of the patients identified as potential readmission should be sick (meaning that the precision of the model should be at least 50%). However, the client also emphasized that the number of wrongful discharges (i.e. the number of false negatives) should be minimized. Note that every model will have a tradeoff between these two metrics, and that satisfying the minimal requirement on precision can lead to significantly increasing the rate of false negatives.
Regarding the no-discrimination criteria, the client requested that the readmission rate should not vary more than 10 percentage points between sub-groups, and less than 5 percentage points between medical specialties.

2. Dataset analysis

2.1 General analysis

The client provided the dataset containing 81412 encounters with the patients dating back to 2012. Initial cleaning of the data was performed as detailed in the annex 5.1. There were 34 variables in the dataset. The variables can be divided into four groups:

1) **General patient information** (gender, race, age, weight, blood type, whether patient has prosthesis, patient's vaccination status)
2) **General information on current encounter** (codes representing type of admission and discharge, and source of admission; medical specialty of physician in charge; insurance status of the patient; time spent in hospital)
3) **Information on tests and procedures performed and diagnosis set during current encounter** (number of tests and procedures; number of medications administered; hemoglobin level; glucose serum level; A1c test result; whether diuretics, insulin or any other diabetic medication was prescribed; whether there was change in medication prescribed; number of diagnosis; primary, secondary and additional secondary diagnosis)
4) **Information on previous encounters** (number of outpatient, emergency and inpatient visits of the patient in the year preceding the encounter

After initial cleaning of the dataset, the consistency of the data was checked by looking at the variables that should not change for the same patient. These were race, gender and blood type. In total, there were 12647 patient_ids that were repeated, meaning that these patients visited the hospital more than once. Out of these 12647 patients, 10055 (79.5%) patients were recorded to have different blood types on different encounters. This suggests that there might be a problem in the way that the value of this variable is collected and that it is not reliable enough to include in the model. However, since this information is not important for answering business question about potential discrimination, these patients were not removed from the dataset.

Regarding race and gender variables, there were 94 patients (0.7%) for whom the race was different in different encounters (not considering unknown), and 2 patients (0.02%) for whom the gender was different (not considering unknown). Since one of the business questions was to check if there is discrimination based on race or gender, these patients were removed from the dataset. However, the fact that the proportion of these patients is low suggests that the data on gender and race is in general reliable. There were 81115 encounters left in the dataset after these patients were removed.

Finally, some patients either died during their stay at the hospital, or were sent to the hospice following the discharge. Since these patients are not expected to be readmitted, they were removed from the dataset. More precisely, these were the patients for which the value of the variable discharge_disposition_code were 11, 13, 14, 19, 20 or 21. There were 79194 encounters left in the dataset after these patients were removed. All further analysis, as well as the development of the model was performed on this dataset.

Percentage of missing values, number of unique values and data type for 34 variables in the cleaned dataset can be found in the table 5.1.1. in annex 5.1. To assess the quality of available data, the percentage of missing values was further analyzed (Figure 2.1.1). The variable “weight” was missing in ~97% of encounters, therefore, it was excluded from the further analysis and in development of the model. Two variables that have high rates of missing values are “payer_code”, which represents insurance provider (~40%) and “medical_specialty”, which refers to the specialty of the physician that is in charge of the patient (~49%). Other variables have less than 10% of missing values, with the large majority having less than 5%.

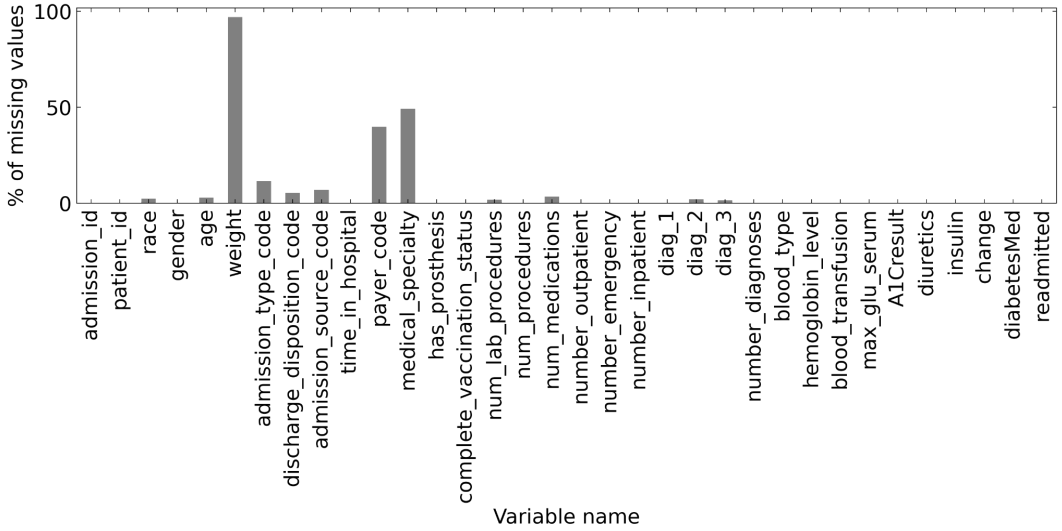

Figure 2.1.1. Percentage of missing values for variables present in the original dataset.

Regarding the distribution of the variables that will be important for business questions, gender was equally distributed (54% female and 46% male), while insurance status was not (only 8% of patients were not insured). Distributions of subgroups in age, race and medical specialty variables were also unequal, as shown in Figure 2.1.2. Regarding age, there was a bias towards older patients, with a low percentage of patients below 30 years. Regarding race, more than 90% of patients were either Caucaisan or African American. There was also high inequality in distribution of medical specialties, with four specialties representing more than 60% of the data (Figure 2.1.2.).

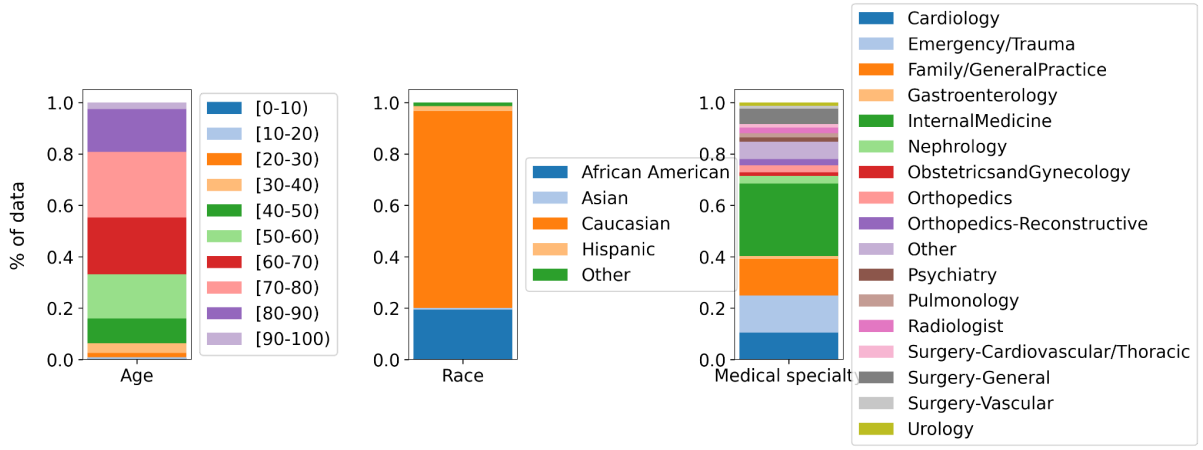

Figure 2.1.2. Distribution of subgroups in age, race and medical specialty variables.

In total, 8992 patients were readmitted in 30 days following the discharge, meaning that the rate of wrongful discharges was 11.4%. This means that the dataset is highly unbalanced. Correlation matrix revealed no correlation between numerical features of the dataset and readmission status (Figure 2.1.3). Furthermore, there were no significant correlations between numerical features.

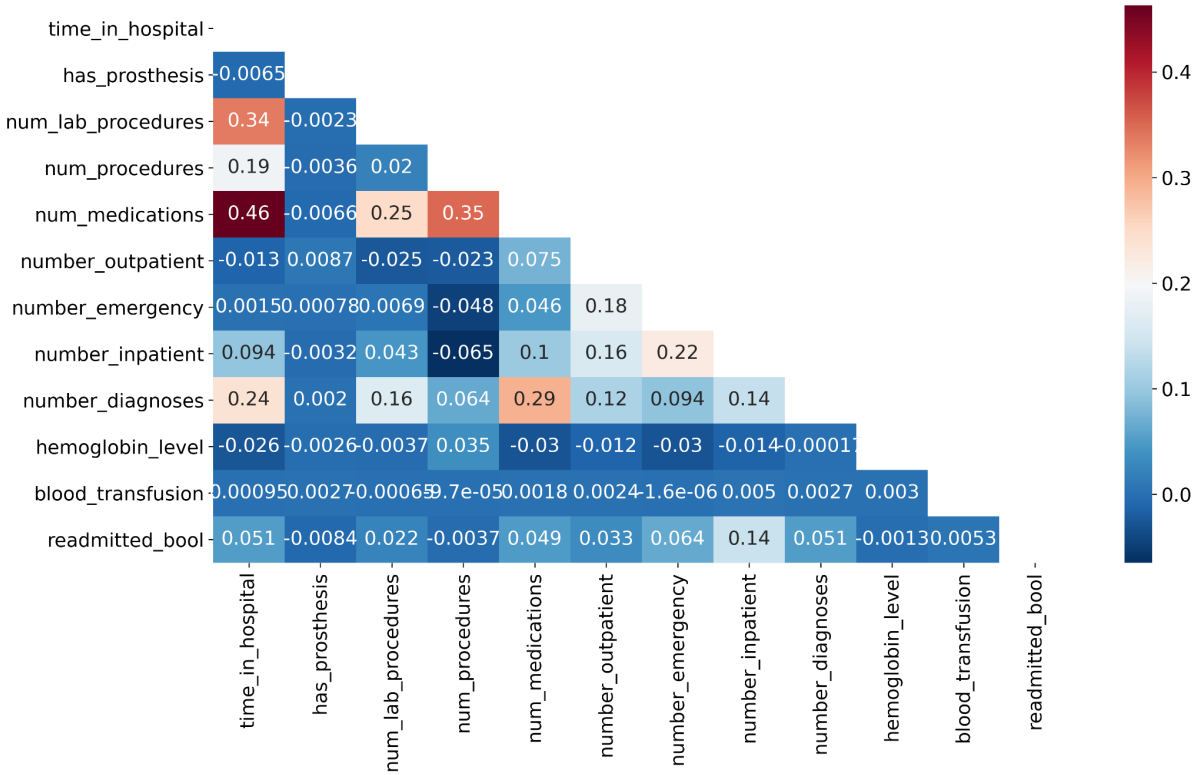

Figure 2.1.3. Correlation matrix between numerical features in the dataset. The numbers represent Spearman's correlation.

2.2 Business questions analysis

The client requested us to analyze whether there is any discrimination based on the age, gender, race or insurance status of the patient. For this, we calculated readmission rates of these sub-groups of patients. We used two criteria to detect potential discrimination. First one is whether the rate of readmission is more than 20% higher than the average rate of readmission. Since the average rate of readmission is 11.4%, the rate that would indicate potential discrimination is ~14%. The second criteria is the difference in readmission rate between sub-groups. According to the client's requirements, the readmission rate should not vary more than 10 percentage points between sub-groups.

We did not find the readmission rate higher than 14% for any gender, race or insurance status subgroup (Figure 2.2.1). Likewise, readmission rates did not vary more than 10 percentage points between sub-groups. For race, the highest difference in readmission rates was 1.7 percentage points (between African American and Other). For gender, the difference between readmission rates was only 0.22 percentage points. Finally, the difference between readmission rates between insured and uninsured patients was 1 percentage point. However, considering that ~40% of the patients do not have the information on their insurance status, this result might not be as reliable.

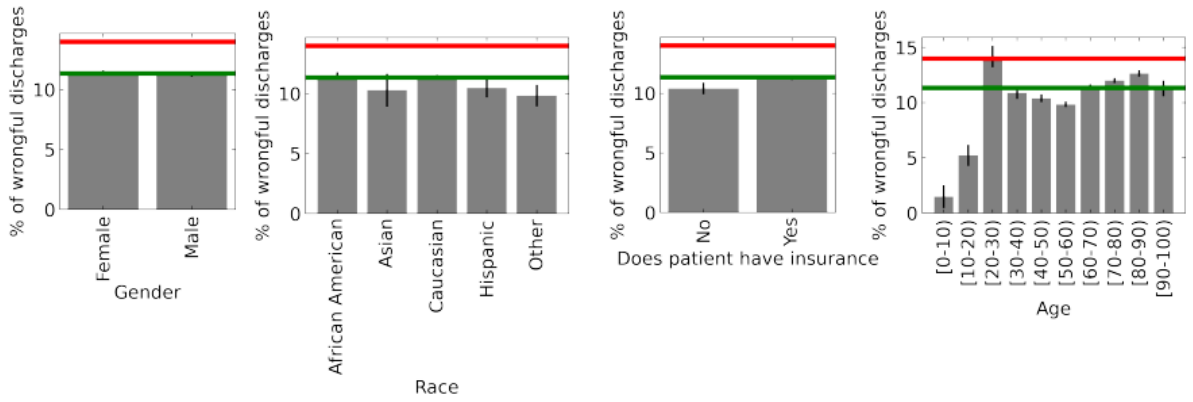

Figure 2.2.1. Percentage of wrongful discharges based on race, gender, insurance status and age of the patient. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error.

Regarding age, the percentage of wrongful discharges was 14.2% for the patients between 20 and 30 years old (Figure 2.2.1.). This indicates that this group might be discriminated against. Furthermore, the difference between readmission rate of this group and the group with the lowest readmission rate (0-10 years old) was 12.7 percentage points. However, since the group of 0-10 years had only 2 wrongful discharges, the significance of this difference is not very high.

The client also requested us to assess whether physicians of any medical specialty might be responsible for higher than average rates of wrongful discharges. Since there are 69 unique medical specialties in the data, and some of them have very low number of encounters and readmissions, we only considered the specialties that had at least 5 wrongful discharges. We again used the threshold of 14% to detect potentially problematic specialties (Figure 2.2.2.). Additionally, according to the client's requirements, the readmission rate should not vary more than 5 percentage points between different specialties.
There were 9 specialties that had a percentage of wrongful discharges higher than 14% (see Table 5.2.1 in annex). The best performing specialty was Obstetrics and Gynecology with a readmission rate of 5.1%. 18 out of 29 specialties considered had a readmission rate that was more than 5 percentage points higher than this value (see Table 5.2.1 in annex). However, since the reference rate of 5.1% is much lower than the average readmission rate, this does not pose a base for concern. Note that 49% of the encounters did not have information on medical specialty, which might have skewed our analysis. Additional analysis of medical specialty data, stratified based on age, gender, race and patient's insurance status can be found in annex 5.2. This analysis indicates that the medical specialties that are contributing the most to the discrimination of the patients between 20 and 30 years old are Emergency/Trauma and Nephrology.

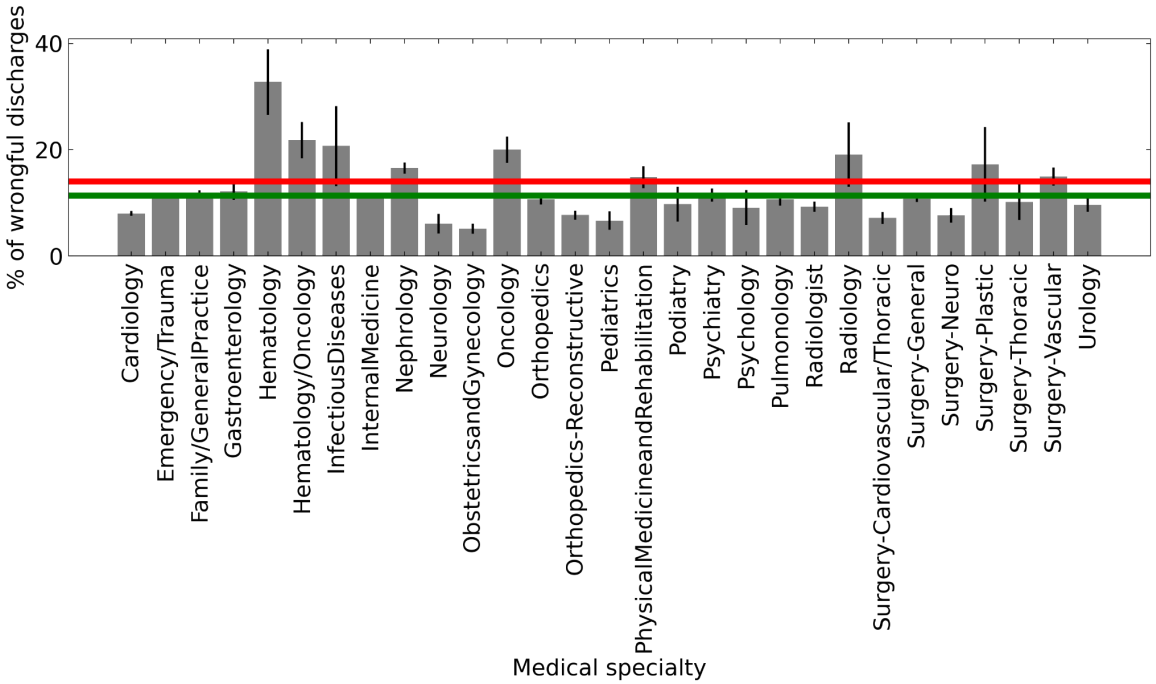

Figure 2.2.2. Percentage of wrongful discharges based on the medical specialty of the physician. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error.

Finally the client requested us to assess whether any admission sources might be responsible for higher than average rates of wrongful discharges. We used the same criteria for filtering data and assess the potential discrimination as for medical specialty (Figure 2.2.3.). The only potentially problematic source was HMO Referral with 14.4% of wrongful discharges and 5.5 percentage points above the best performing source. However, this was very marginal.

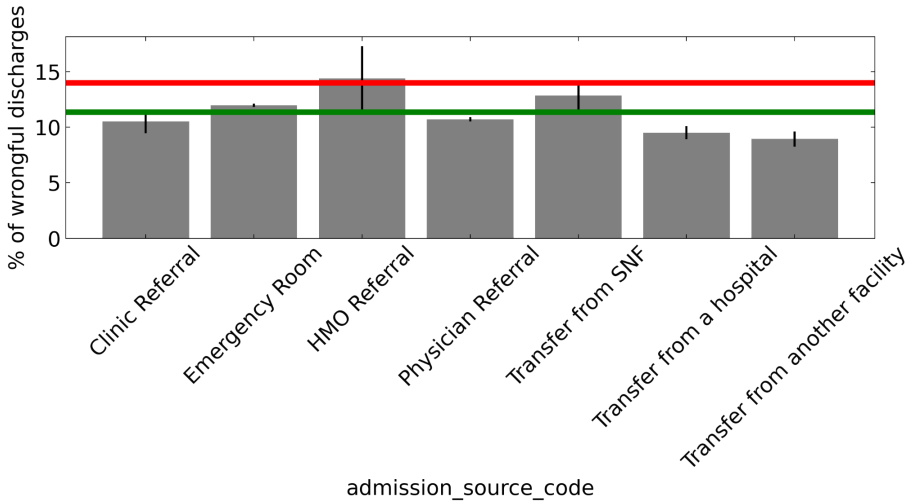

Figure 2.2.3. Percentage of wrongful discharges based on the admission source. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error.

2.3 Conclusions and Recommendations

Based on the general data analysis done in chapter 2.1, we advise the client to reassess the method of gathering and recording the patient data. We found that ~80% of the patients that visited the hospital more than once had different recordings of blood types on different occasions. This could indicate the problem in the way in which the blood type data is being collected. Even more concerning, this could indicate a more general problem in the data collecting pipeline, meaning that the whole dataset might not be reliable.
Additionally, the information on the weight of the patient was missing in 97% of the cases. We advise the client to collect this data more consistently, as well as to add the data such as patient height. This would allow to test wheteher the physicians are discriminating based on patient's weight or body mass index, which is a known issue in medical field.

We found indications that the patients between 20 and 30 years old are more often wrongfully discharged, which might require further investigation. More importantly, we found a few specialties for which the readmission rate was much higher than average. We advise the client to carry out a detailed investigation about why this is the case.

Finally, the client initially warned that some training and changes of staff and/or policy may impact the data over time. However, since the time stamps of the encounters were not provided in the dataset, we could not have analyzed if this was the case. We advise the client to add the information on the date and time of the encounter in the future.

3. Modeling

3.1 Model expected outcomes overview

Based on preliminary results, we expect a reduction of readmission rate by ~38%. However, the client's requirement that at least 50% of the patients identified as potential readmissions should actually be sick could not be satisfied. Currently, only around 17% of the patients identified as potential readmissions turned out to actually be readmitted. Note that this percentage could be increased, but only at the cost of many more patients being wrongfully discharged.

Ragarding the equality criteria, the model satisfies criteria of no discrimination based on age, gender, race and insurance status of the patient. However, it does not satisfy the requirement that the readmission rates should vary less than 5 percentage points between medical specialties. Nevertheless, the model significantly reduces differences between medical specialties compared to the original dataset.

3.2 Model specifications

There were several changes performed on the dataset before it was used to train the model:

1. To build the target variable, the values of the "readmitted" feature were replaced with True if the original value is "Yes", and False if the original value is "No".
2. Age groups, which were received as ranges, were coded as integers to preserve information about order ([0-10) as 0, [10-20) as 10…)
3. Values of features "admission_type_code", "discharge_disposition_code" and "admission_source_code" were replaced with "Other" if they appear less than 100 times in the dataset.
4. ICD-9 codes were replaced with one of the 18 types of disease they represent (source: https://en.wikipedia.org/wiki/List_of_ICD-9_codes)
5. Features "weight", "payer_code" and "medical_specialty" were not used because of a large number of missing values (section 2.1.). Feature "blood type" was not used because of problems with data consistency explained in section 2.1.

Final list of 27 features that were used to train the model, including the number of unique and missing values, can be found in Table 5.3.1. in annex 5.3.

Dataset was split between train and test set (70% train and 30% test). In the train dataset, all rows that had at least one missing value were dropped. This reduced the dataset by ~10%. Train dataset was undersampled before fitting because the dataset is highly unbalanced (section 2.1.). Numerical features were scaled by removing the mean and scaling to unit variance, while categorical features were encoded using one hot encoding. In the test dataset, missing values of numerical features were imputed using median value.

After considering a few alternatives (see section 3.4.), Random Forest classifier was chosen as the best estimator for the problem. Optimal values of hyperparameters "n_estimators" (the number of trees in the forest) and "max_depth" (the maximum depth of the tree) were found using hyperparameter tuning (see annex 5.3 for details). They were found to be

n_estimators=500 and max_depth=5. All other hyperparameters were kept at their default value[1]. The most important features of the final model are shown in figure 3.2.1.

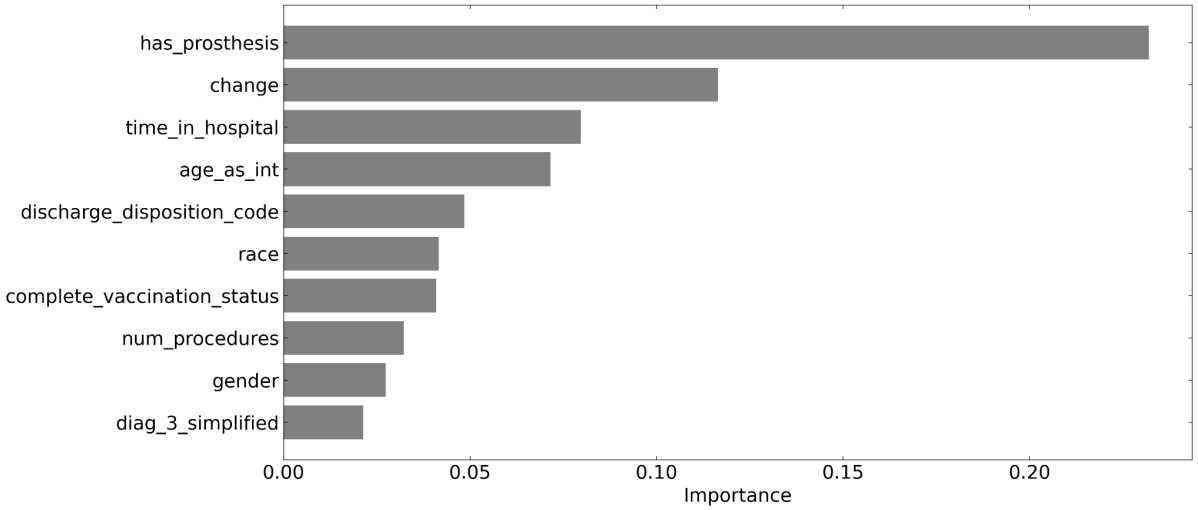

Figure 3.2.1. Ten most important features for the Random Forest prediction model.

3.3 Analysis of expected outcomes based on training set

The values of relevant metrics of the model based on the training set are:

- Readmission rate: 0.071
- F1-score: 0.266
- Recall: 0.655
- Precision: 0.167
- ROC AUC: 0.618

Overall readmission rate was reduced by ~38% (from 0.114 to 0.071). The precision level is ~17%, meaning that only around 17% of patients that were identified as potential readmissions were readmitted. This is much lower than requested 50%, and could be increased by increasing the threshold for labeling the observation as True. However, it would significantly decrease the recall score, which is the percentage of readmitted patients that were correctly identified as potential readmissions (Figure 3.3.1.). In particular, setting the threshold to the value that satisfies the client's criteria of 50% precision rate, would lead to almost all patients being discharged. This would also reduce the ROC AUC score to ~50%, meaning that the model is completely incapable of distinguishing between the classes. Ultimately, the decision on the optimal threshold will depend on the cost of keeping the patient for no reason, versus the cost of having the patient readmitted. Although in the currently deployed model the threshold is set to 50%, this could be easily changed if requested by the client. A good criteria for setting threshold could be maximization of F1-score, which combines both precision and recall.

[1] The default parameters of Random Forest Classifier can be found at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

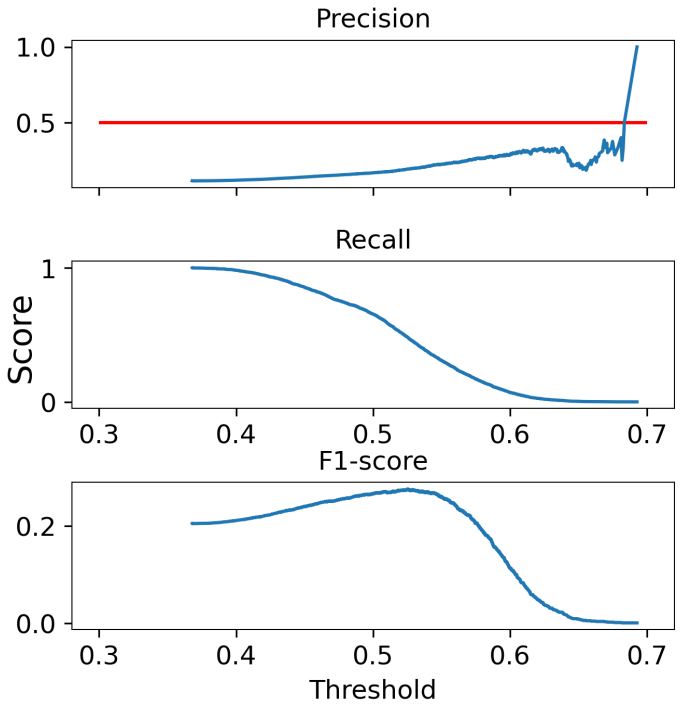

Figure 3.3.1. Change in precision, recall and f1-score with changing threshold for labeling the observation as True. Red line in the topmost plot indicates a precision score of 0.5, which was the client's requirement for minimal precision.

Regarding the non-discrimination criteria, the model satisfies criteria of no discrimination based on age, gender, race and insurance status of the patient (the readmission rates do not differ more than 10 percentage points between sub-groups). However, it does not satisfy the criteria that the readmission rates vary less than 5 percentage points between medical specialties. Namely, the difference between medical specialties with highest and lowest readmission rates is around 10% (Table 3.3.1.). However, the model managed to reduce the original bias, since in the original dataset this difference was close to 30%. Additionally, the number of unique values of medical specialty features is much larger than the number of unique values in gender, age, race and insurance status features. Because of this, the observed differences between admission rates of best and worst performing specialties are more likely to be due to chance. To test this, we first performed pairwise comparison of proportions, which compares each proportion with all the others. We corrected for the fact that we are doing multiple comparisons using the Holm–Bonferroni method. This analysis did not yield a statistically significant ($p<0.05$) difference between any of the subgroups. However, the client set stricter criteria on differences between medical specialties than on differences between other features. We would advise to adjust these criteria according to the number of unique values.

Table 3.3.1. Differences of admission rates between the groups with highest and lowest rates. Note that only medical specialties that occur at least 50 times in the test data were considered.

Feature	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
race	Other	9.8	Hispanic	1.59	8.21
age	[60-70)	8.05	[30-40)	2.65	5.4
gender	Male	6.84	Female	6.28	0.56
isInsured	True	6.63	False	6.08	0.55
medical_specialty	Surgery-Cardiovascular/Thoracic	12.5	Psychiatry	2.56	9.94

3.4 Alternatives considered

Five classifiers were considered: Logistic regression classifier, Random forest classifier, Gradient boosting classifier and Support vector classifier. Grid search was performed to find the best hyperparameters for each of the classifiers, with F1 score used as the metric (see table 5.3.2 in annex for details on hyperparameter tuning). All classifiers performed very similarly (Table 3.4.1).

Table 3.4.1. Comparison of performance of five classifiers tested.

Classifier	Readmission rate	F1 score	Recall	Precision	ROC AUC
Logistic Regression	0.077	0.272	0.571	0.179	0.616
Decision Tree	0.071	0.255	0.695	0.156	0.606
Random Forest	0.071	0.266	0.655	0.167	0.618
Gradient Boosting	0.073	0.268	0.628	0.170	0.617
Support Vector	0.075	0.268	0.602	0.173	0.615

Because all classifiers performed similarly, the classifier to be used was decided based on the performance regarding non-discrimination criterial. All five classifiers satisfied the criteria of no discrimination based on age, gender and insurance status of the patient (Tables 5.3.3., 5.3.4. and 5.3.5. in the annex 5.3.). Decision Tree, Random Forest and Support vector classifiers satisfied the criteria of no discrimination based on race (Table 5.3.6. in annex 5.3). None of the classifiers satisfied the criteria that the readmission rates vary less than 5 percentage points between medical specialties. However, Random Forest performed better than Decision Tree and Support Vector classifiers (Table 5.3.7. in annex 5.3). Finally, Random Forest offers a good compromise on speed of training, being much faster than Gradient Boosting or Support Vector.

3.5 Known issues and risks

Any model can be as good as is the data that is used to train it. With this in mind, it is important to note that the original dataset had some inconsistencies, which might reflect deeper problems in the dataset (see chapter 2).
Some features that were present in the dataset were not used in the current version of the model because of the high amount of missing values. However, these features might be important. If the values of these features are collected more consistently in the future, the model could be retrained taking them into account. Additionally, some missing values are being imputed which might negatively influence the performance of the model.

Expected outcomes of the model presented here are based on the performance on the test dataset. However, if there are significant changes of the data over time, the performance of the model might change. If this happens, the retraining of the model will be necessary.

4. Model Deployment

4.1 Deployment specifications

The application is developed using the Flask framework and deployed on Heroku platform under the Free pricing plan using Heroku Git deployment method. The Heroku Postgre database is used to store requests and predictions. The database is integrated with the app using pewee object–relational mapping. There are four columns in the database: observation_id (unique identifier of observation), observation (request sent to the /predict endpoint), proba (predicted probability of readmission) and true_class (actual readmission status).

The url of the application is https://capstone-amilas.herokuapp.com/.
There are two endpoints:

1) **predict,** to which the client can send the request for new predictions. Prediction will be “Yes” if the patient is predicted to be readmitted in the following 30 days, or “No” if not.
2) **update**, to which the client can send the information on whether the patient has been readmitted in the following 30 days.

The structure of expected requests and responses can be found in Table 5.4.1 in annex 5.4.
There are few criteria that have to be satisfied in order to successfully process the request:
For /predict endpoint:

- admission_id is either integer or convertible to integer (e.g. “101”, “101.0”, 101.0)
- admission_id is not already present in the database
- all of the columns that were present in the original dataset are present in the request

For /update endpoint,

- admission_id is present in the database
- the value of the “readmitted” field is either “Yes”, if the patient has been readmitted, or “No” if not.

If any of the above criteria is not satisfied, the response code 400 and description of the error will be returned. If an unexpected error happens during handling of the request, the response code 500 will be returned.

Additionally, the values of some features will be converted in order to satisfy requirements of the model:

- All of the categorical features used in the model will be converted into one of the known categories using the steps that are described in annex 5.1. and chapter 3.2. The conversion is not case sensitive.
- The values of booleans will be set as True if the entered value is either True (bool), “True” (string) or 1 (either float, integer or string). Otherwise, it will be set as False.
- Numerical categories that represent the number of something (e.g. number of diagnoses, number of procedures) will be converted to integers.

- Time in hospital cannot be more than 40000 and less than 0 days. Value of hemoglobin level cannot be more than 100 and less than 0. All other numerical values have to be between 0 and 1000.

If it is not possible to convert the value of the features to the format that is expected by the model, or if the numerical value is not inside the expected range, it will be set to NaN.

4.2 Known issues and risks

Heroku platform has some limitations that could cause problems with the service, the extensive documentation on this can be found at https://devcenter.heroku.com/articles/limits. Importantly, there are limits on the number and size of dynos, and number of processes that can exist in a dyno. This could lead to problems in performance if the service is to be extensively used. Additionally, the calls to the API are limited to a rate of at most 4500 calls per hour.
Regarding the limitation of the database, under the current pricing plan, the database has a limit of 10 000 entries. If the service is to be used continually, this limit will be reached very quickly. Additionally, the data is not encrypted, which is concerning since the patient's medical data is handled. Considering these limitations, the upgrade of the pricing plan could be necessary.

In the current version of the service, all fields that do not satisfy criteria will be set as NaN and the prediction will be given. This means that even the requests that have no valuable information on the patient's status will yield prediction. However, these will be completely unreliable. We suggest that the response could be updated so that it gives warning about fields that are missing data.
Another issue is the current handling of booleans. Since the model requires these fields to be of bool type, it cannot be set as NaN. Therefore, the value that is sent is going to be considered True only if it is given as True (bool), "True" (string) or 1 (either float, integer or string). All other cases will be considered False. In the following implementations the model could be updated so that it does not require these fields to be of bool type.
Finally, the client is currently getting only categorical "Yes" or "No" as the answer to the question if the patient is likely to be readmitted. However, it might be more useful to get the probability of readmission. In this way, the practitioner could use the combination of the model's output and their own judgment to assess whether the patient should be discharged. This feature can be easily implemented into the service if the client wishes to do so.

5. Annexes

5.1 Dataset technical analysis

DATA CLEANING STEPS:

1) All ‘?’ values were replaced with NaN

2) Cleaning of columns:

FEATURE: race
CLEANING PROCESS:
values ["EURO", "European", "WHITE", "White"] were replaced with "Caucasian"
values ["AfricanAmerican", "Afro American", "AFRICANAMERICAN", "Black"] were replaced with "African American")
Value “Latino” was replaced with “Hispanic”
UNIQUE VALUES AFTER CLEANING: ['Caucasian' 'African American' 'Other' nan 'Hispanic' 'Asian']

FEATURE: gender
CLEANING PROCESS: Value “Unknown/Invalid replaced with NaN
UNIQUE VALUES AFTER CLEANING: ['Male' 'Female' nan]

FEATURE: admission_type_code
CLEANING PROCESS: values 5 (No Available), 6 (NULL) and 8 (Not Mapped) were replaced with NaN; data type was set as string
UNIQUE VALUES AFTER CLEANING: [['3.0' '2.0' '1.0' 'nan' '4.0' '7.0']

FEATURE: discharge_disposition_code
CLEANING PROCESS:
values 18 (NULL), 25 (Not Mapped) and 26 (Unknown/Invalid) were replaced with NaN; data type was set as string
UNIQUE VALUES AFTER CLEANING: ['1.0' '22.0' '3.0' '11.0' 'nan' '4.0' '6.0' '14.0' '2.0' '23.0' '5.0'
'13.0' '15.0' '7.0' '16.0' '24.0' '28.0' '8.0' '9.0' '27.0' '17.0' '10.0' '19.0' '12.0']

FEATURE: admission_source_code
CLEANING PROCESS: values 9 (Not Available), 15 (Not Available), 17 (NULL), 20 (Not Mapped) and 21 (Unknown/Invalid) were replaced with NaN; data type was set as string
UNIQUE VALUES AFTER CLEANING: ['1.0' '7.0' 'nan' '2.0' '4.0' '6.0' '5.0' '3.0' '13.0' '22.0' '14.0'
'8.0' '10.0' '11.0' '25.0']

FEATURE: medical_specialty
CLEANING PROCESS: value “PhysicianNotFound” was replaced with NaN
UNIQUE VALUES AFTER CLEANING: [nan 'Emergency/Trauma' 'InternalMedicine' 'Family/GeneralPractice'
'Radiologist' 'Orthopedics' 'Cardiology'

'PhysicalMedicineandRehabilitation' 'Orthopedics-Reconstructive'
'Psychiatry' 'Surgery-Thoracic' 'Surgery-Vascular' 'Hematology/Oncology'
'Osteopath' 'ObstetricsandGynecology' 'Pediatrics' 'Nephrology'
'Otolaryngology' 'Urology' 'Surgery-General' 'Anesthesiology-Pediatric'
'Surgery-Cardiovascular/Thoracic' 'Gastroenterology' 'Pulmonology'
'Oncology' 'Podiatry' 'Surgery-Neuro' 'Pediatrics-Endocrinology'
'Neurology' 'Obstetrics' 'Endocrinology' 'Hospitalist' 'Pathology'
'Surgery-Pediatric' 'Radiology' 'Surgery-Plastic'
'Pediatrics-CriticalCare' 'Psychology' 'Psychiatry-Child/Adolescent'
'Ophthalmology' 'Surgery-Cardiovascular' 'Pediatrics-Pulmonology'
'DCPTEAM' 'Obsterics&Gynecology-GynecologicOnco' 'Surgeon'
'Pediatrics-EmergencyMedicine' 'Hematology' 'AllergyandImmunology'
'SurgicalSpecialty' 'OutreachServices' 'Rheumatology' 'Dentistry'
'Gynecology' 'InfectiousDiseases' 'Anesthesiology' 'Surgery-Colon&Rectal'
'Pediatrics-Neurology' 'Perinatology' 'Neurophysiology'
'Psychiatry-Addictive' 'Dermatology' 'Proctology' 'Cardiology-Pediatric'
'Speech' 'Resident' 'Endocrinology-Metabolism' 'Surgery-Maxillofacial'
'Pediatrics-AllergyandImmunology' 'Pediatrics-Hematology-Oncology'
'SportsMedicine']

FEATURE: max_glu_serum
CLEANING PROCESS:
Value “NONE” was replaced with “None”
Value “NORM” was replaced with “Norm”
UNIQUE VALUES AFTER CLEANING: ['None' '>300' 'Norm' '>200']

Table 5.1.1. Percentage of missing values, number of unique values and data type for 34 variables in the cleaned dataset.

	% of missing values	Number of unique values	Data type
admission_id	0.000000	79194	int64
patient_id	0.000000	58692	int64
race	2.265323	5	object
gender	0.002525	2	object
age	2.860065	10	object
weight	96.917696	9	object
admission_type_code	11.508448	5	float64
discharge_disposition_code	5.381721	19	float64
admission_source_code	6.907089	14	float64
time_in_hospital	0.000000	14	int64
payer_code	39.783317	17	object
medical_specialty	49.042857	69	object
has_prosthesis	0.000000	2	bool
complete_vaccination_status	0.000000	3	object
num_lab_procedures	1.833472	115	float64
num_procedures	0.000000	7	int64
num_medications	3.290653	73	float64
number_outpatient	0.000000	39	int64
number_emergency	0.000000	29	int64
number_inpatient	0.000000	21	int64
diag_1	0.018941	699	object
diag_2	1.982473	722	object
diag_3	1.425613	756	object
number_diagnoses	0.000000	16	int64
blood_type	0.000000	8	object
hemoglobin_level	0.000000	77	float64
blood_transfusion	0.000000	2	bool
max_glu_serum	0.000000	4	object
A1Cresult	0.000000	4	object
diuretics	0.000000	2	object
insulin	0.000000	2	object
change	0.000000	2	object
diabetesMed	0.000000	2	object
readmitted	0.000000	2	object

5.2 Business questions technical support

Table 5.2.1. Medical specialties that had percentage of readmission rate higher than 10.1%

	Medical specialty	% of readmissions
0	Hematology	32.76
1	Hematology/Oncology	21.77
2	InfectiousDiseases	20.69
3	Oncology	20.00
4	Radiology	19.05
5	Surgery-Plastic	17.24
6	Nephrology	16.55
7	Surgery-Vascular	14.92
8	PhysicalMedicineandRehabilitation	14.81
9	Gastroenterology	12.13
10	Family/GeneralPractice	11.88
11	Psychiatry	11.44
12	Emergency/Trauma	11.31
13	InternalMedicine	11.29
14	Surgery-General	10.75
15	Pulmonology	10.64
16	Orthopedics	10.60
17	Surgery-Thoracic	10.13

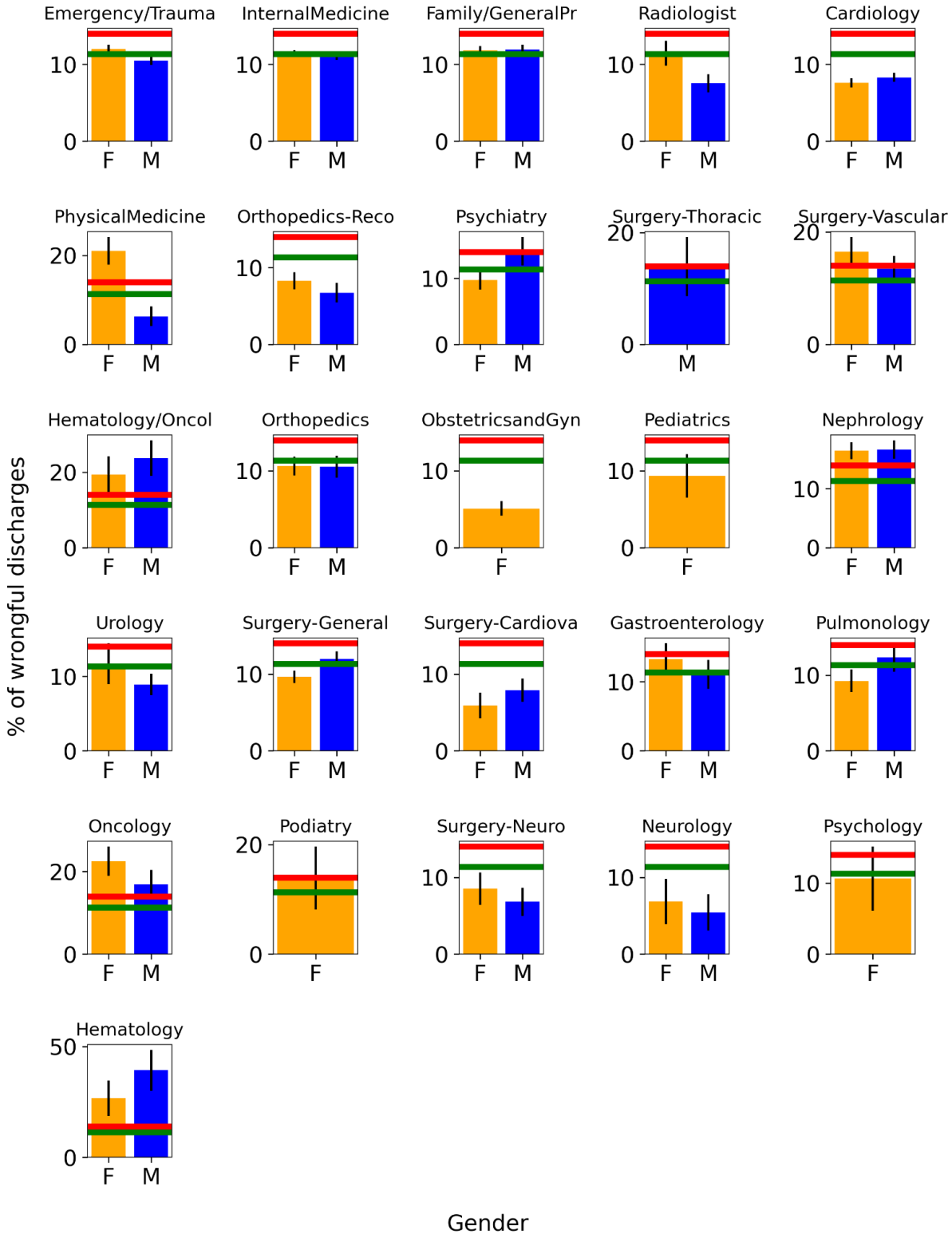

Figure 5.2.1. Percentage of wrongful discharges based on the medical specialty of the physician, stratified by gender. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error. F=Female, M=Male.

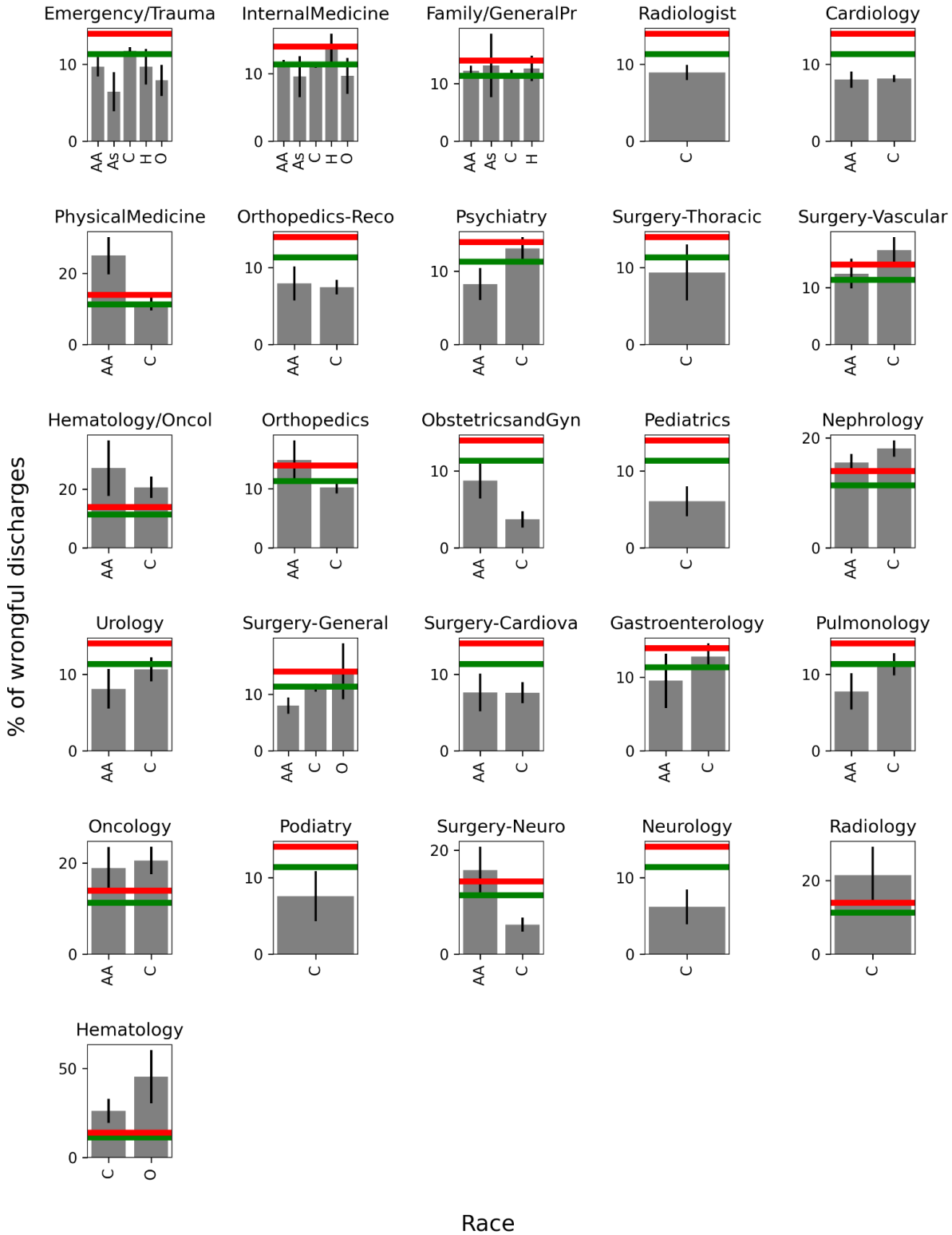

Figure 5.2.2. Percentage of wrongful discharges based on the medical specialty of the physician, stratified by race. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error. AA = African American, As = Asian, C = Caucasian, H = Hispanic, O = Other.

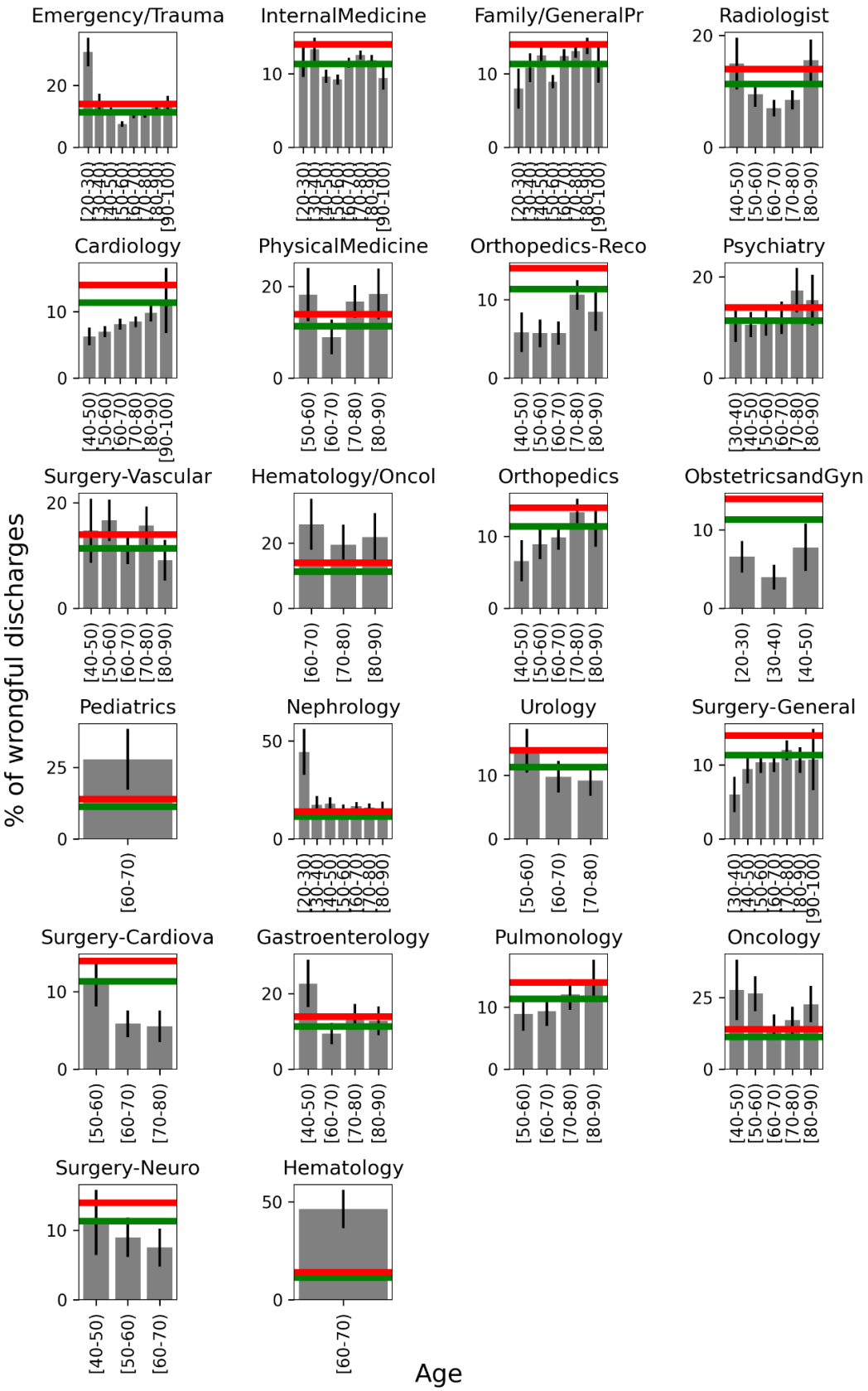

Figure 5.2.3. Percentage of wrongful discharges based on the medical specialty of the physician, stratified by age. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error.

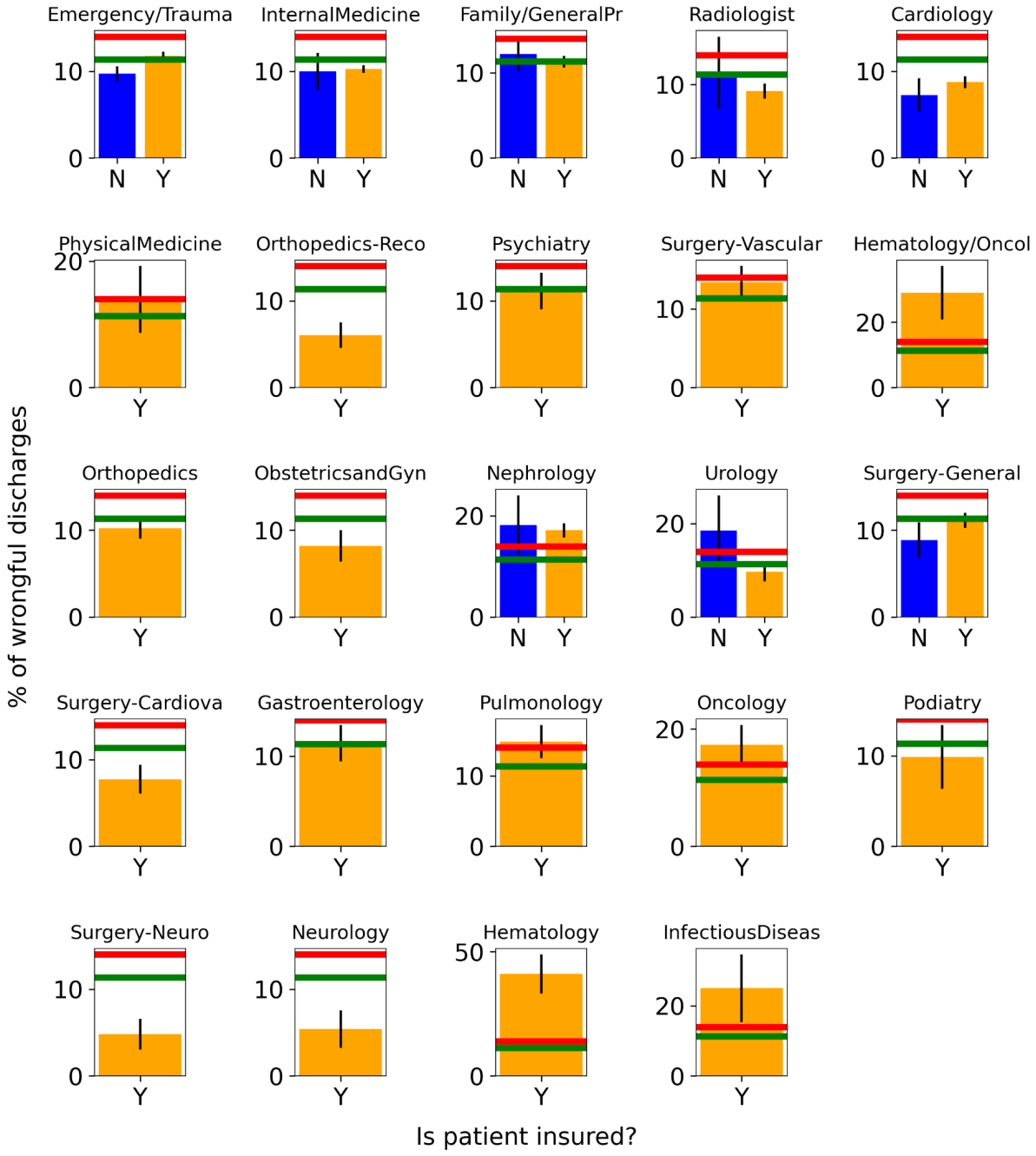

Figure 5.2.4. Percentage of wrongful discharges based on the medical specialty of the physician, stratified by patient's insurance status. Green horizontal line represents the average percentage of wrongful discharges, while the red line represents the percentage of discharge that is 20% higher than the average (14%). Error bars represent standard error. Y = Yes, N = No.

5.3 Model technical analysis

Table 5.3.1. Percentage of missing values, number of unique values and data type for 27 variables that were used to train the model.

	% of missing values	Number of unique values	Data type
race	2.265323	5	object
gender	0.002525	2	object
admission_type_code	11.508448	4	object
discharge_disposition_code	5.381721	11	object
admission_source_code	6.907089	8	object
time_in_hospital	0.000000	14	int64
has_prosthesis	0.000000	2	bool
complete_vaccination_status	0.000000	3	object
num_lab_procedures	1.833472	115	float64
num_procedures	0.000000	7	int64
num_medications	3.290653	73	float64
number_outpatient	0.000000	39	int64
number_emergency	0.000000	29	int64
number_inpatient	0.000000	21	int64
number_diagnoses	0.000000	16	int64
hemoglobin_level	0.000000	77	float64
blood_transfusion	0.000000	2	bool
max_glu_serum	0.000000	4	object
A1Cresult	0.000000	4	object
diuretics	0.000000	2	object
insulin	0.000000	2	object
change	0.000000	2	object
diabetesMed	0.000000	2	object
age_as_int	2.860065	10	float64
diag_1_simplified	0.018941	17	object
diag_2_simplified	1.982473	17	object
diag_3_simplified	1.425613	17	object

Table 5.3.2. Hyperparameter space used to find the optimal hyperparameters for tested classifiers.

Classifier	Hyperparameters tested	Best hyperparameters
Logistic Regression	{"C": [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100], "penalty": ['none', 'l1', 'l2', 'elasticnet']}	{'C': 0.01, 'penalty': 'l2'}
Decision Tree	{"max_depth": [1, 3, 5, 7, 9]}	{'max_depth': 5}
Random Forest	{'n_estimators': [100, 200, 500, 1000], 'max_depth' : [1, 3, 5, 7, 9]}	{'max_depth': 5, 'n_estimators': 500}
Gradient Boosting	{'n_estimators': [100, 200, 500, 1000], 'max_depth' : [1, 3, 5, 7, 9]}	{'max_depth': 3, 'n_estimators': 100}
Support Vector	{'kernel': ["linear", "poly", "rbf", "sigmoid"], 'C': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100]}	{'C': 1, 'kernel': 'rbf'}

Table 5.3.3. Performance of five classifiers based on criteria of no discrimination by age. The client's requirement was that there is no more than 10 percentage point difference between subgroups.

Classifier	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
LogisticRegression	[90-100)	9.23	[30-40)	3.82	5.41
DecisionTreeClassifier	[60-70)	8.35	[30-40)	2.08	6.27
RandomForestClassifier	[60-70)	8.05	[30-40)	2.65	5.4
GradientBoostingClassifier	[60-70)	8.42	[30-40)	2.7	5.72
SVC	[90-100)	8.62	[30-40)	3.23	5.39

Table 5.3.4. Performance of five classifiers based on criteria of no discrimination by gender. The client's requirement was that there is no more than 10 percentage point difference between subgroups.

Classifier	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
LogisticRegression	Male	7.26	Female	6.9	0.36
DecisionTreeClassifier	Female	6.64	Male	6.3	0.34
RandomForestClassifier	Male	6.84	Female	6.28	0.56
GradientBoostingClassifier	Female	6.59	Male	6.58	0.01
SVC	Male	7.16	Female	6.4	0.76

Table 5.3.5. Performance of five classifiers based on criteria of no discrimination by insurance status. The client's requirement was that there is no more than 10 percentage point difference between subgroups. True = patient is insured, False = patient is not insured.

Classifier	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
LogisticRegression	True	7.3	False	5.48	1.82
DecisionTreeClassifier	True	6.57	False	5.87	0.7
RandomForestClassifier	True	6.63	False	6.08	0.55
GradientBoostingClassifier	True	6.64	False	6.2	0.44
SVC	True	6.87	False	5.98	0.89

Table 5.3.6. Performance of five classifiers based on criteria of no discrimination by race. The client's requirement was that there is no more than 10 percentage point difference between subgroups.

Classifier	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
LogisticRegression	Other	12.07	Hispanic	0.0	12.07
DecisionTreeClassifier	Other	8.89	Hispanic	0.0	8.89
RandomForestClassifier	Other	9.8	Hispanic	1.59	8.21
GradientBoostingClassifier	Other	10.91	Hispanic	0.0	10.91
SVC	African American	9.09	Hispanic	0.0	9.09

Table 5.3.7. Performance of five classifiers based on criteria of no discrimination by medical specialty. The client's requirement was that there is no more than 5 percentage point difference between medical specialties. Note that only medical specialties that occur at least 50 times in the training data were considered.

.

Classifier	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
LogisticRegression	Pulmonology	12.2	Psychiatry	2.38	9.82
DecisionTreeClassifier	Gastroenterology	15.0	Psychiatry	4.17	10.83
RandomForestClassifier	Surgery-Cardiovascular/Thoracic	12.5	Psychiatry	2.56	9.94
GradientBoostingClassifier	Nephrology	11.96	Psychiatry	2.33	9.63
SVC	Pulmonology	13.16	Psychiatry	2.5	10.66

5.4 Model deployment technical support

Table 5.4.1. Examples of expected input and output of predict and update endpoints

Endpoint	Expected payload (input)	Expected response (output)

POST /predict	{'admission_id': 82679,
	'patient_id': 81317898,
	'race': 'Caucasian',
	'gender': 'Male',
	'age': None,
	'weight': '?',
	'admission_type_code': 3.0,
	'discharge_disposition_code': 3.0,
	'admission_source_code': 1,
	'time_in_hospital': 7,
	'payer_code': 'MC',
	'medical_specialty': 'Family/GeneralPractice',
	'has_prosthesis': False,
	'complete_vaccination_status': 'Complete',
	'num_lab_procedures': 44.0,
	'num_procedures': 1,
	'num_medications': 13.0,
	'number_outpatient': 0,
	'number_emergency': 0,
	'number_inpatient': 2,
	'diag_1': '585',
	'diag_2': '491',
	'diag_3': '276',
	'number_diagnoses': 9,
	'blood_type': 'O+',
	'hemoglobin_level': 15.6,
	'blood_transfusion': False,
	'max_glu_serum': 'None',
	'A1Cresult': 'None',
	'diuretics': 'No',
	'insulin': 'Yes',
	'change': 'Ch',
	'diabetesMed': 'Yes'}
		{
		"readmitted": "Yes"
		}

POST /update	{
	"admission_id": 82679,
	"readmitted": "No"
	}
		{
		"admission_id": 82679,
		"actual_readmitted": "No",
		"predicted_readmitted": "Yes"
		}