LISBON DATA SCIENCE ACADEMY

ANALYSIS AND PREDICTION OF HOSPITAL DISCHARGES

ASSESSMENT OF THE PERFORMANCE OF THE MACHINE-LEARNING MODEL

Prepared for:
Hazel and Bazel Hospital

Prepared by:
Ana Milas
Data scientist
Awkward Problem Solutions

6 of March 2022

Table Of Contents

1. Business Conclusions

1.1 Summary

The client, Hazel and Bazel Hosipital, requested us to develop a model that would predict if the patient will be readmitted in the 30 days following the discharge. Regarding the performance of the model, the client requested that at least 50% of the patients identified as potential readmission should be sick (i.e. the precision of the model should be at least 50%). However, the client also emphasized that the rate of identifying the healthy patients as potential readmissions should be minimized (i.e. the number of false negatives should be minimized).

The client also had some requirements that should prevent the discrimination of the patient's based on gender, ethnicity, age, insurance status or medical specialties of the practicioner. In particular, readmission rate should not vary more than 10 percentage points between sub-groups of gender, race, age and insurance status, and less than 5 percentage points between medical specialties.

The model was developed using the initial data provided by the client, which contained details on around 80 000 unique encounters. This model was delivered to the client in the form of REST API. The client then used this API to get predictions for around 10 000 new encounters, as well as to report on whether the patient was readmitted after 30 days.

In this report, we detail the model's performance on the new data taking into consideration client's requirements and expected performance. Furthermore, we suggest how the performance and the deployment of the model could be further improved.

In summary, the model performed as expected based on the analyses reported in the Report 1. Both regarding the precision of the model, and no-discrimination criteria. However not all of the client's requirements were met. In particular, we found it hard to satisfy the requirement on precision of the model, while keeping the number of false negatives low. Only 16.5% of the patients identified as potential readmissions were eventually readmitted. However, we were able to reduce the rate of wrongful discharges by ~35% (from 11.3% to 7.3%).
Regarding the no-discrimination criteria, the model satisfied requirements of no-discrimination based on age, gender, race or insurance status. However, it did not satisfy the criteria that the readmission rates vary less than 5 percentage points between medical specialties. The highest difference between two medical specialties was 14.3% ("Psychiatry" and "Surgery-Cardiovascular/Thoracic")

2. Results Analysis

2.1 Model Performance

In total, 9696 requests with unique values of “admission_id” were sent to the /predict endpoint. The actual labels for all of these observations were also sent to /update endpoint, which allowed us to compare expected and actual performance of the model. The comparison of performance measured by metrics considered in the Report 1 is shown in table 2.1.1.

Table 2.1.1. Expected and actual performance on the metrics highlighted in the Report 1. Difference represents the difference between the second and the first column.

Metric	Expected	Actual	Difference	Percentage difference
Readmission rate	0.071	0.073	0.002	2.82%
F1-score	0.266	0.262	-0.004	-1.50%
Recall	0.655	0.638	-0.017	-2.60%
Precision	0.167	0.165	-0.002	-1.20%
ROC AUC	0.618	0.613	-0.005	-0.81%

The performance of the model on the new data was very close to expected performance on all metrics considered. This shows that the model presented in the Report 1 is generalizable and can be used on new data. Additionally, this indicates that the relationship between features used to build the model and the readmission rates did not significantly change in the new dataset.

As highlighted in the Report 1 (see page 11 and Figure 3.3.1 in Report 1), the precision of the model could be improved if the threshold for labeling the observation as True is increased. Although the threshold was not changed from the default value of 0.5, we wanted to test how well we can predict the change in precision, recall and F1-score with changing threshold (Figure 2.1.1). This can be useful for making the decision on the threshold for future iterations of the model.The dependence of recall and F1-score on threshold did not change considerably between the two datasets. However, there was a difference in the precision of the model on original and new dataset when the threshold was increased. Additionally, the precision of the model on both datasets showed very noisy behavior for higher thresholds. This shows that increasing the threshold level above ~0.55 is not a good idea because it could lead to very unpredictable performance of the model.

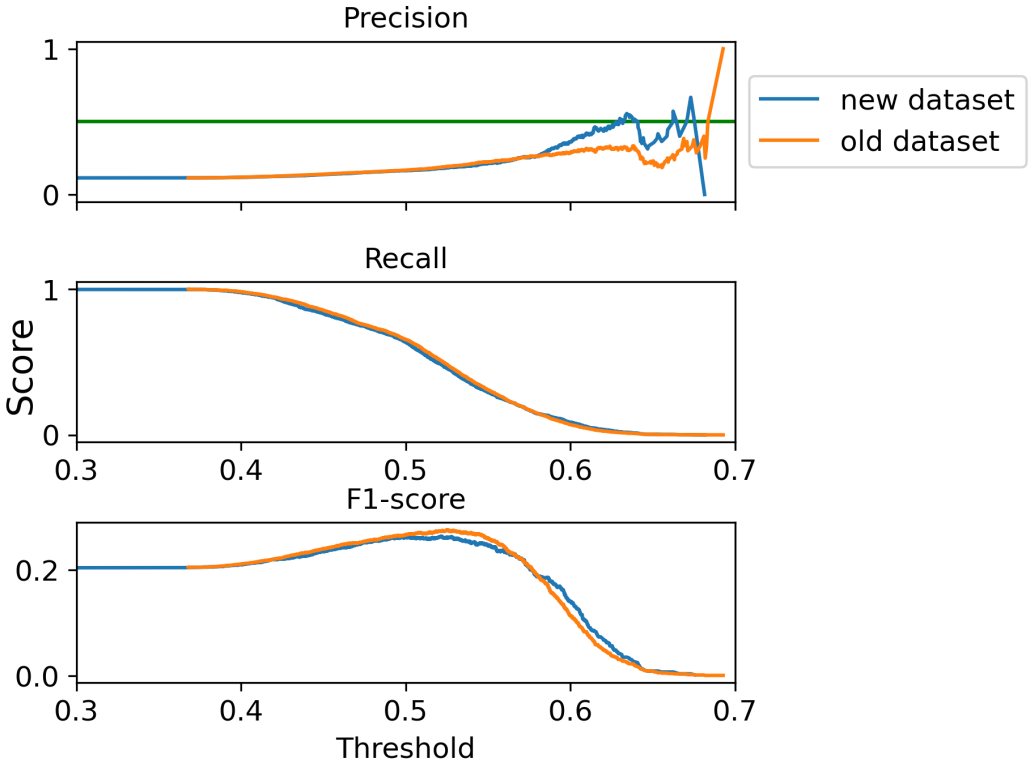

Figure 2.1.1. Change in precision, recall and f1-score with changing threshold for labeling the observation as True. Green line in the topmost plot indicates a precision score of 0.5, which was the client's requirement for minimal precision.

2.2 Success on requirements

The client had two requests regarding the performance of the model. First, the model should have a precision of at least 50%. Second, the number of wrongful discharges should be minimized. The current model has precision of 16.5% on the new data, and therefore does not satisfy the first requirement. However, we were able to reduce the rate of wrongful discharges by ~35% (from 11.3% to 7.3%). The performance on both of these requirements is as expected based on the original data (see section 2.1.).
As explained in chapter 3 of Report 1 and in chapter 2.1. of this report, we did not decide to increase the precision by increasing the threshold for labeling the observation as True. This is something that can be done upon the clients requests in future iterations of the model. However, increasing the threshold will negatively influence the performance on the second requirement. More specifically, increasing the precision of the model to 50% would require us to increase the threshold so much that the rate of wrongful discharges would be 10.9%.

Regarding the no-discrimination criteria, the client requested that the readmission rate should not vary more than 10 percentage points between race, age, gender and insurance status sub-groups, and less than 5 percentage points between medical specialties.
When assessing the model's performance on these criteria in Report 1, we only considered the subgroups that appear in the test set at least 50 times. In this way, we tried to avoid reporting very high or very low rates of readmission that are the consequence of low representation of the subgroup, and not the consequence of discrimination.
When using this approach on the new data, the model satisfies criteria of no discrimination based on race, age, gender and insurance status of the patient. 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 14.3% (Table 2.2.1).

Table 2.2.1 Differences of readmission rates between the groups with highest and lowest rates. Note that only subgroups 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	African American	8.41	Hispanic	4.55	3.86
age	[90-100)	12.7	[20-30)	4.29	8.41
gender	Female	7.76	Male	7.28	0.48
isInsured	True	7.57	False	6.75	0.82
medical_specialty	Psychiatry	16.67	Surgery-Cardiovascular/Thoracic	2.38	14.29

The performance of the model on discrimination criteria is in agreement with the expectations reported in Report 1 (Table 3.3.1 in Report 1). Interestingly, the medical specialty with the highest rate of readmission on the test set was Surgery-Cardiovascular/Thoracic (12.50%, Report 1), while this specialty had the lowest rate of readmission in the new dataset (2.38%, Table 2.2.1). Contrary, Psychiatry had the lowest rate of readmission in the test set (2.56%, Report 1), and the highest rate in the new dataset (16.67%, Table 2.2.1). This change is probably due to change in the actual readmission rates of these two specialties in original and new datasets (Figure 2.2.1). Since the medical specialty feature was not included in the building of the model, this change should not

directly influence the performance of the model. However, there could be features included in the model that correlate with medical specialty, and therefore changed in similar ways between the original and the new dataset. This observation indicates that retraining of the model on the new dataset might be necessary.

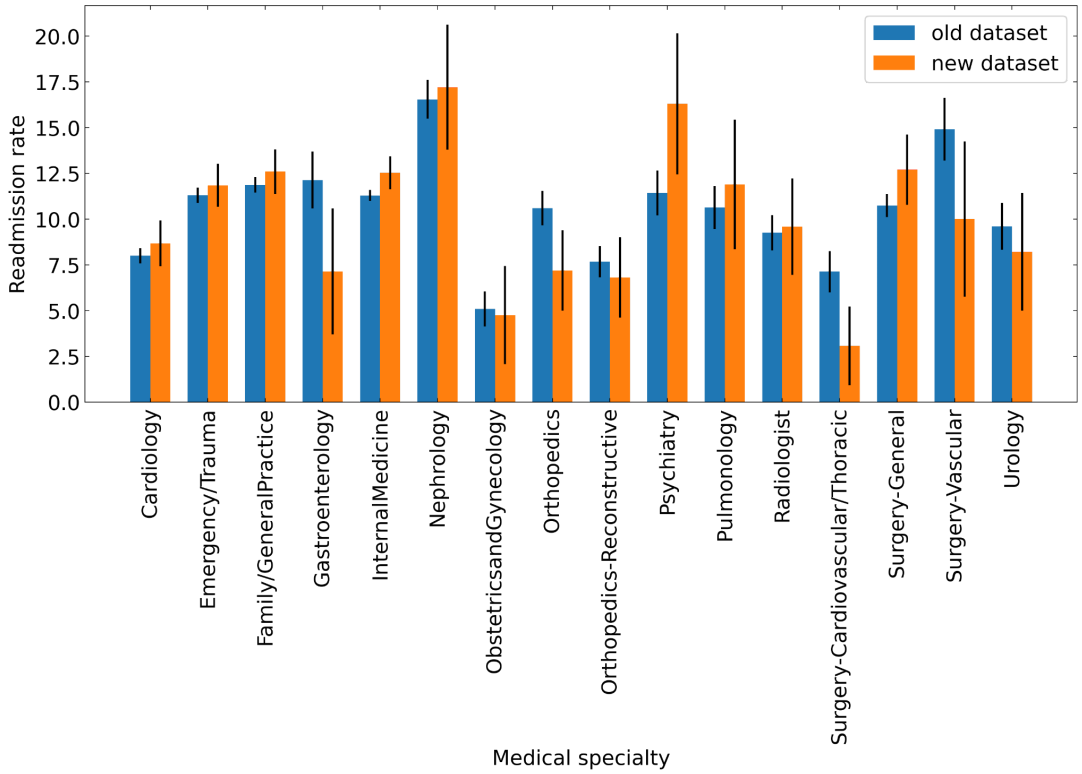

Figure 2.2.1 Readmission rate based on the medical specialty of the physician in original and new dataset. Error bars represent standard error. Only specialties that appear at least 50 times in the dataset are considered. Note the change in readmission rates between the old and new datasets for “Psychiatry” and “Surgery-Cardiovascular/Thoracic”.

Finally, the table 2.2.2 shows the differences in readmission rates if no limit on the minimal number of occurrences of the subgroup in the dataset is set. In this case, the difference in readmissions of the age groups [90-100) and [0-10) is 12.7%, which is above the maximum difference criteria. However, the age group [0-10) is represented with only 10 examples. The maximal difference between medical specialties is 50%. However, the number of encounters where the medical specialty of the practitioner was “Infectious Diseases” or “Anesthesiology” was only four and one, respectively.

Table 2.2.2 Differences of readmission rates between the groups with highest and lowest rates. No filtering based on the sample size was performed

Feature	Group with highest rate	Highest rate	Group with lowest rate	Lowest rate	Difference
race	Asian	8.57	Hispanic	4.55	4.02
age	[90-100)	12.7	[0-10)	0.0	12.7
gender	Female	7.76	Male	7.28	0.48
isInsured	True	7.57	False	6.75	0.82
medical_specialty	InfectiousDiseases	50.0	Anesthesiology	0.0	50.0

2.3 Population Analysis

The new dataset contained 9696 observations. Distributions of variables relevant to the client's business questions (gender, race, age, insurance status and medical specialty) did not change between the new and the original dataset (Figure 2.3.1).

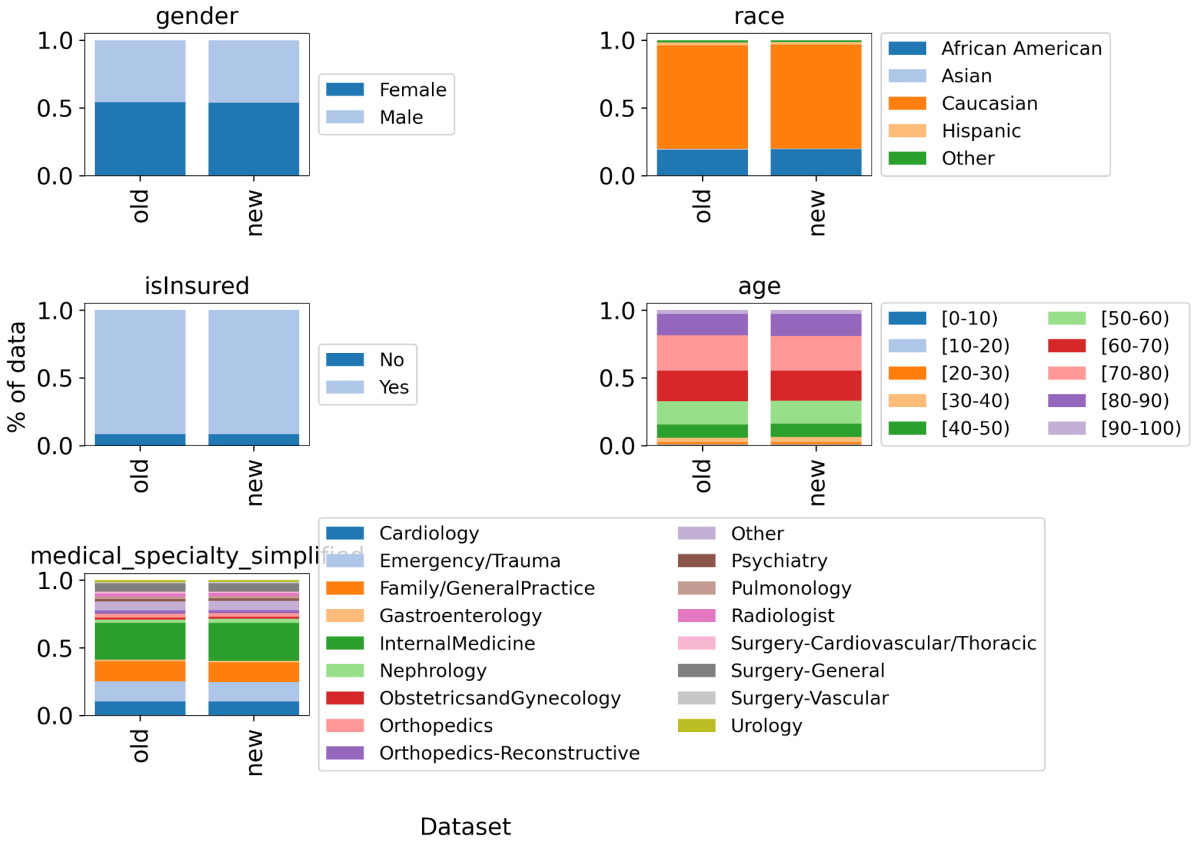

Figure 2.3.1. Distribution of subgroups in gender, race, age, insurance status and medical specialty variables in original and new dataset. "Other" subgroup in the medical specialty data contains all specialties that represent less than 1% od the data in the original dataset.

The average readmission rates did not significantly change between the original (11.4%) and the new dataset (11.3%). Regarding the readmission rates between subgroups for gender, race, age and insurance status, the differences were mostly apparent between the subgroups that were less represented in the dataset (e.g. Hispanic and Other race subgroup; patients without insurance, very young patients) (Figure 2.3.2). Similar differences were present between less represented medical specialties (see Figure 2.2.1 in the previous section). This indicates that the observed differences are more likely due to statistical fluctuations, than due to actual changes in treatments of the subgroups between the two datasets.

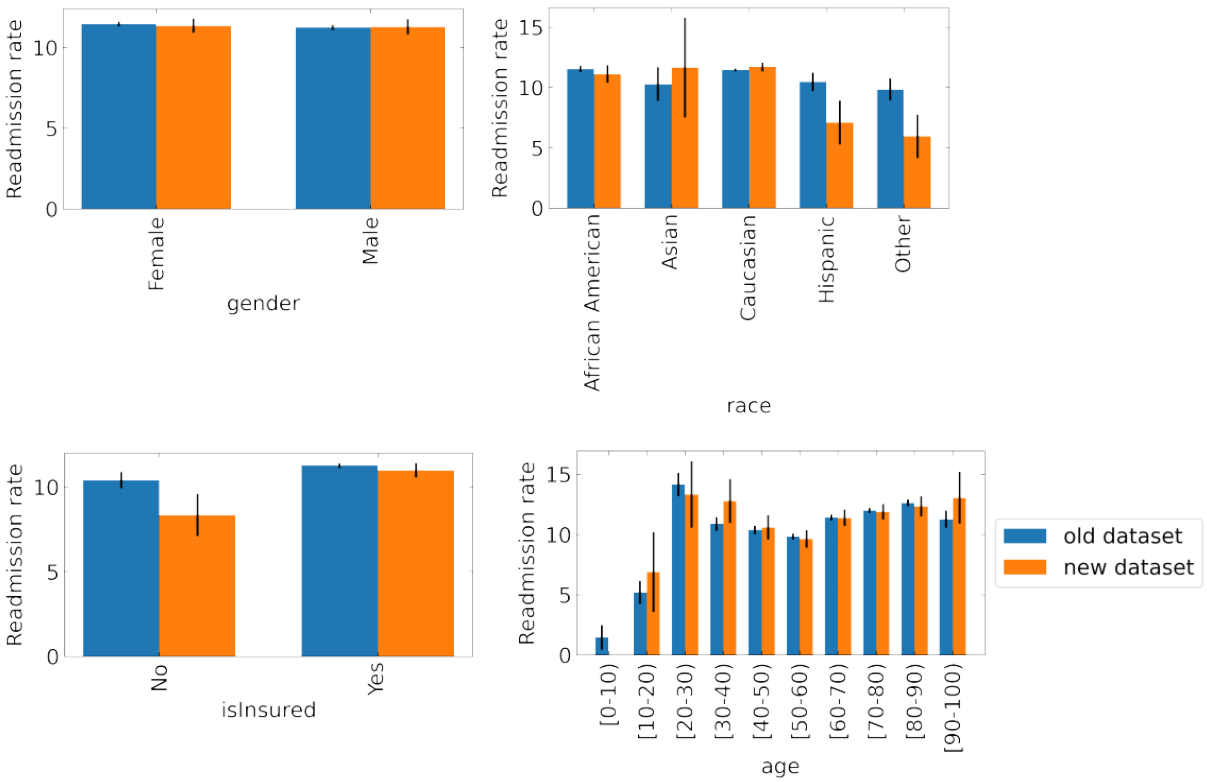

Figure 2.3.2. Readmission rates in percentages based on race, gender, insurance status and age of the patient in original and new dataset. Error bars represent standard error. Note that the readmission rate for the age group [0-10) was 0.

The distributions of features that were used to build the model were very similar in the two datasets, as shown in Figure 2.3.3 for numerical features and Figure 2.3.4 for categorical features.

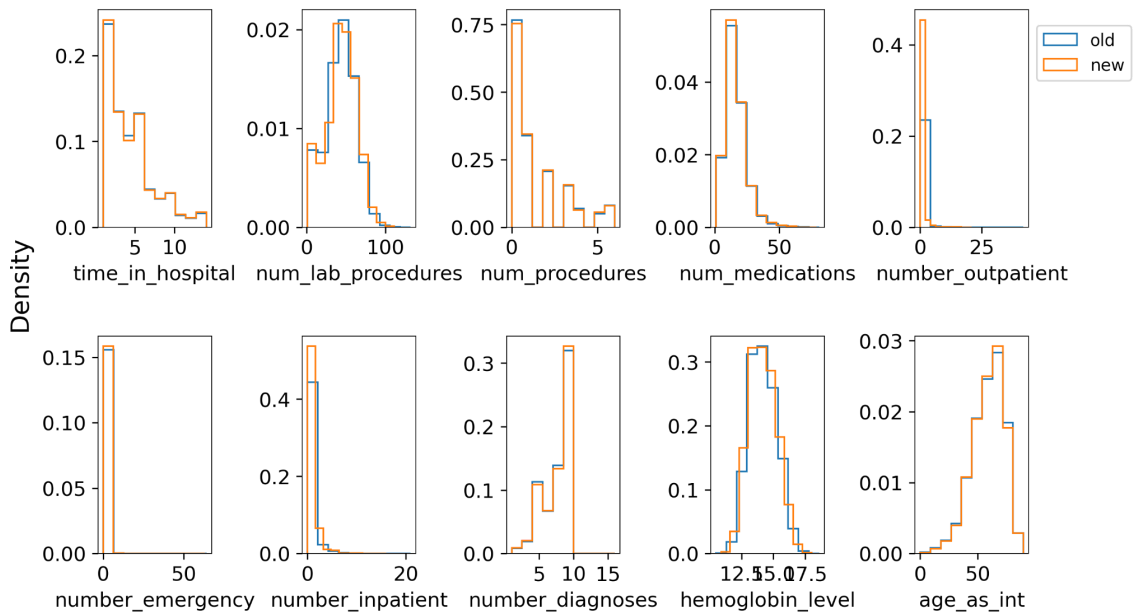

Figure 2.3.3 Distribution of numerical variables that were used to build the model. Blue: original dataset, orange: new data.

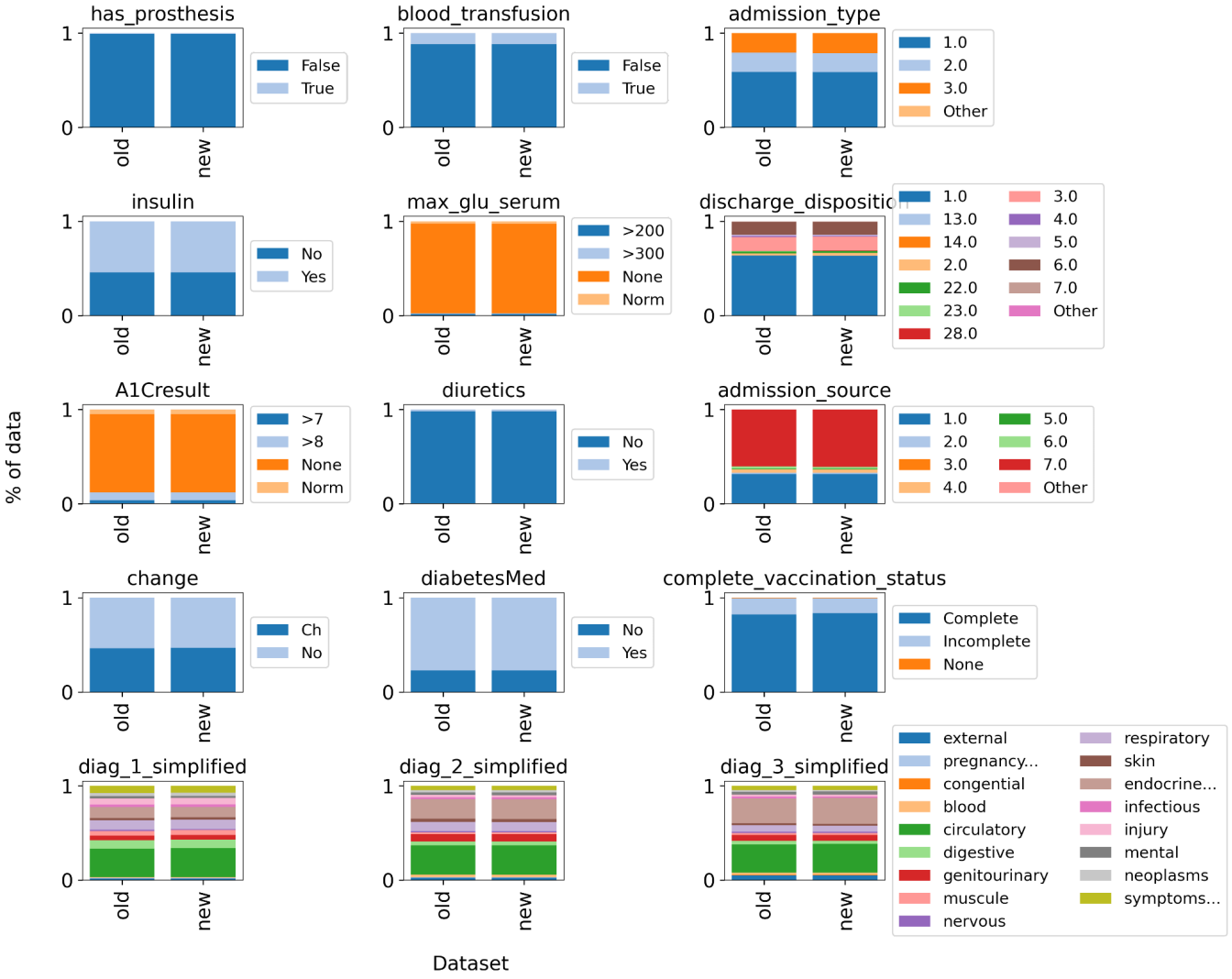

Figure 2.3.4 Distribution of categorical variables that were used to build the model in original and new dataset.

3. Next Steps

3.1 Next Steps

The model does not satisfy the client's requirement that at least 50% of the patients identified as potential readmissions are actually sick. That is, the precision of the model is not up to the client's requirement. As already discussed in chapter 2, the precision of the model could be increased by increasing the threshold for labeling the observations as True. However, this was not done in the current version of the model, because it would lead to significantly increasing the readmission rate. Nevertheless, this could be done upon client's requirement in the future iterations of the model.

There were several steps that were already tried in the attempt to increase the performance of the model. For example, performing feature engineering by taking squares or square roots of the numerical features, or reducing the number of features by removing the features with low variance. These approaches did not improve the performance. However, we did notice that removing some features did not decrease the performance either. Therefore, it might be useful to remove these features from the model in the future to make it less computationally heavy.

As reported in Report 1, the model was selected by testing several classifiers and carefully tuning for the hyperparameters. However, all models tested performed similarly. Finally, the current model was retrained on all available data (both original and new encounters) but it did not improve the performance. The fact that additional data did not improve the model indicates that the features that were used do not have enough information for accurate prediction. Therefore, acquiring new features describing the encounters might be necessary to improve the model. Some features that might be useful are the patient's weight (which was present here, but the data was missing for >95% of the patients) or more details on the treatments and test results of the patient.

In general, the model performed well on non discrimination criteria. It showed some variability in classifying the less represented subgroups. However, this is expected since wrongly classified examples will be more strongly reflected on the readmission rate when the sample size is small. Therefore, we advise the client to take this into account when assessing the possible discrimination.
In addition, the differences between subgroups were larger for the groups that have larger numbers of unique values (e.g. medical specialty). This is also expected, since there are higher chances to obtain big differences by chance when there are more subgroups. However, the client set stricter criteria on differences between medical specialties than on differences between other features, which have less unique values. We would advise to adjust these criteria according to the number of unique values.

4. Deployment Issues

4.1 Re-deployment

The model was not redeployed for several reasons. First, there were no serious issues in the functioning of the API (see next section for details). Second, the performance of the model was as expected (section 2.1. and 2.2.), and distribution of the data was very similar between the new and the original dataset (section 2.3.). Finally, all the attempts to improve the model were unsuccessful (see section 3.1. and the rest of this section for details).

We tried to improve the performance of the model by adding new data to the training set. However, the retrained model performed very similarly to the original model, even showing slight decrease in the performance (Table 4.1.1).

Table 4.1.1. Performance of the original model and the model retrained using additional data. Difference represents the difference between the second and the first column. Red font indicates that the retrained model performed worse than the original model.

	Original	Retrained	Difference	Percentage difference
Readmission rate	0.071	0.074	0.003	4.22%
F1-score	0.266	0.263	-0.003	-1.13%
Recall	0.655	0.619	-0.036	-5.49%
Precision	0.167	0.167	0	0%
ROC AUC	0.618	0.613	-0.005	-0.81%

The original model did not satisfy the client's fairness requirement based on the medical specialty of the practitioner (see Table 2.2.1 in this report and Table 3.3.1 in Report 1). Therefore, potential better performance of the retrained model on this criteria could warrant redeployment. However, the retrained model again performed worse than the original (Table 4.1.2).

Table 4.1.2. Performance of the original and retrained model 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.

	Model	
	Original	**Retrained**
Group with highest rate	Surgery-Cardiovascular /Thoracic	Pulmonology
Highest rate (%)	12.5	20.63
Group with lowest rate	Psychiatry	Obstetrics and Gynecology
Lowest rate (%)	2.56	3.85
Difference between highest and lowest rate	9.94	16.76

4.2 Unexpected problems

In general, there were no serious issues connected to deployment of the model. The application did not crash or have a downtime during the period when the requests were sent by the client.

The observations that were saved to the database did not have unexpected values of the categorical features, and the numerical features were inside the expected ranges. However, there were requests for which the admission_id already existed in the database. For these requests, the prediction was made and returned to the client together with the error message. However, only the original request was saved to the database. This could be problematic if the new request was sent because the previous request had wrong or insufficient data. This could be solved by developing a more interactive API. In addition to sending the warning, the app could also request the user to specify which request contains correct data.
Other problems with the requests were not noticed. However, since the Heroku only stores the last 1500 lines of log history, some of these were maybe missed.

In the current version of the service, all fields of the request that do not satisfy the validation criteria will be set as NaN. This means that the prediction will be sent even for the requests with missing or invalid data in all or most fields. However, this issue did not occur in this round of requests. At most, the requests had 4 missing values of the features relevant for the model.

Finally, the database limit of 10000 rows was almost reached, since there were 9696 requests sent.

4.3 Learnings and Future Improvements

Overall, this project was a great learning experience because it allowed us to develop several different skills. First, data analysis and visualization were used to answer some of the client's business questions, and to explore the dataset provided by the client. Next, this dataset was used to develop and test a machine learning model to predict premature discharges of the patients. Finally, the model was deployed and delivered to the client in a form of REST API which can be used to make new predictions. Apart from this, developing the project from scratch was a valuable experience in project management, since we had to organize our time and efforts in order to meet all the deadlines.

Possible improvements of the model performance are discussed in detail in section 3.1. Most important observation is that addition of the new data did not improve the model (section 3.1. and 4.1). This indicates that currently available features do not contain enough information to make correct predictions. Therefore, future efforts should concentrate on feature engineering or collection of the new features describing the encounters (e.g. patient's weight, test results, treatments).

There are also some improvements to be made in deployment of the model. As already mentioned, limits that Heroku imposes on the log history make it challenging to monitor the potential problems with the app and requests. This could be improved by using packages for logging in Python (e.g. loguru). Future improvements could also include developing more interactive API. This would be particularly useful when the requests with repeated admission_id are sent (section 4.2). Although there were no problems with the data quality in this round of requests, it could be useful to send a warning message to the client if the request has lots of missing or invalid values.