Group 27

Theo Jagodits 21416074
Grant Liu 93027462
Neal Sharma 38568337

UCI Diabetes Data Set

CS 273a Final Report

Theodore Jagodits, Neal Sharma, Grant Liu

Introduction

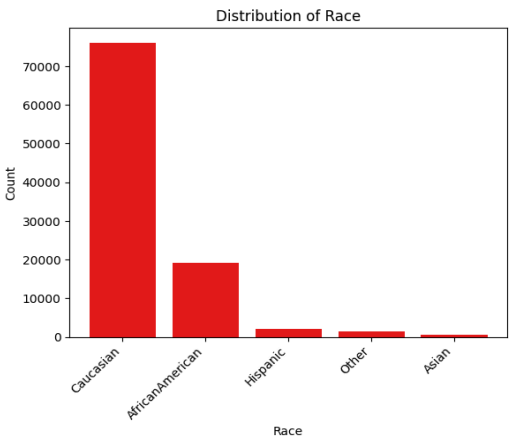
In our concluding project report, we opted for the UCI diabetes hospital dataset. Numerous individuals grappling with diabetes may face challenges in receiving adequate treatment, potentially necessitating readmission for further care. This not only incurs expenses for both patients and hospitals but also signifies a lapse in initial care, contributing to heightened morbidity and mortality rates among patients. The objective of utilizing this dataset for prediction is to determine the likelihood of patient readmission and identify the contributing factors to such occurrences. We will go through the data set and plot various features to get familiar with them as well as perform statistical analysis to determine which ones are important. Subsequently, we will perform feature selection and then finally run two different machine learning models on the data.

Dataset Description

The dataset comprises tabular data containing 101,766 rows and 47 feature columns. It encapsulates a decade of patient information gathered from 130 US hospitals and integrated delivery networks. Patient stays ranged up to 14 days, and their readmission status was monitored within 30 days, after 30 days, or not at all following discharge.

The feature columns consist of various types of medical data. The first part is patient data such as age, race, weight, etc. The second part of the features are about the hospital visit and associated variables such as length of stay, how many times the patient was readmitted that year, and the type of doctor. The last part of the features is about the types of tests and treatments administered that are common for diabetes.

Feature Inspection

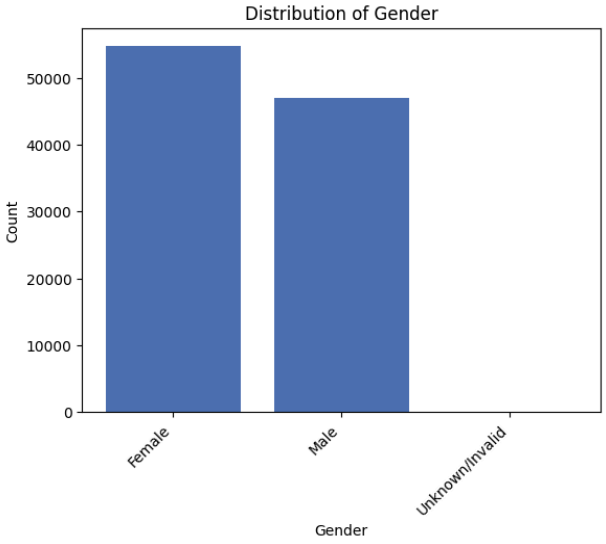


The first part of the medical data is about the patient and is as follows: race, gender, age, and weight. As we can see from the graph, most patients are predominantly Caucasian with the next highest race being African American. There are very few Hispanic, Asian and other races represented in the data set. In fact, Caucasians account for 75% of the data. This might have

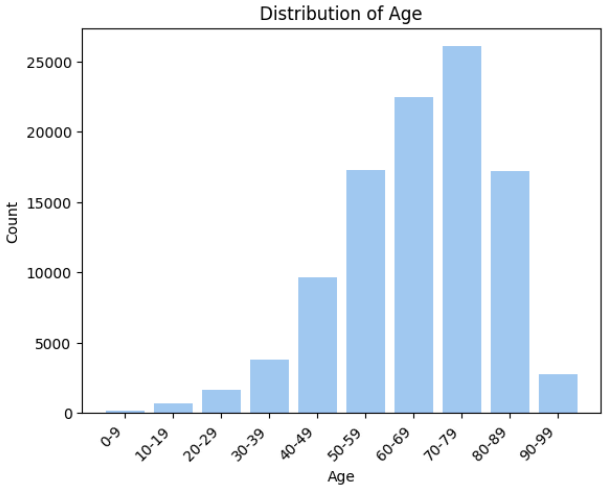
Theo Jagodits 21416074
Grant Liu 93027462
Neal Sharma 38568337

an impact on the bias of certain medications since it is known that different races respond differently to medications.

The next feature is gender, which is fairly even, with 53.7% female and 46.2% male. This is slightly surprising since other studies mention that there are usually more males than females that have diabetes. However, it is also common for females to have greater complications with diabetes which could explain why they have more hospital admissions. There are three instances of Unknown/Invalid gender.



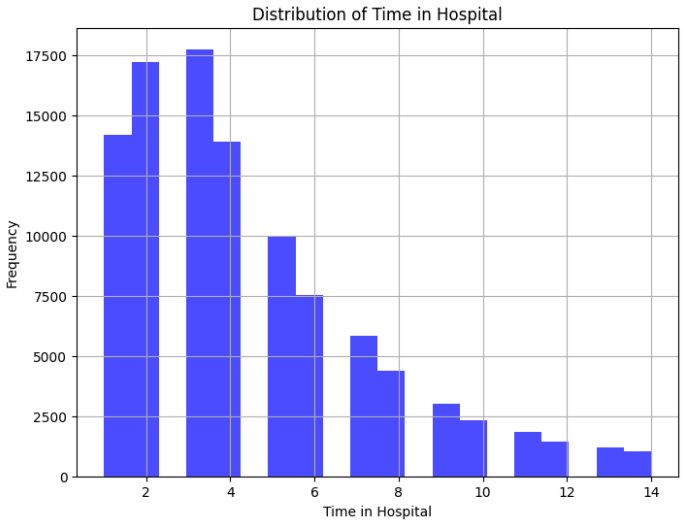
The age distribution is skewed towards the right with older people from age 50-89 being the majority of patients. This is not out of the ordinary, this differs slightly from other statistics with most people with diabetes being 45-64.



Finally in the patient section is weight which is a categorical value in this data set. However, over 95% of patient weight is missing so we will omit this data and graph for brevity since we will remove this feature later on.

The section on features is about the details of the patient visit. They are patient ids, diagnosis, doctor specialties, and etc. Again for sake of brevity we will omit most of these features for analysis and attach graphs and tables at the end of this report. The interesting features are as follows: number of lab procedures, time in hospital, number of medications, and number of diagnoses.

The number of lab procedures is defined as the number of lab tests performed during the patient stay. The median number of lab procedures is 44 with the maximum of 132 and minimum of one lab procedure. The minimum is presumably a A1C test which is common for diabetes patients.



The time in hospital seems like a key indicator if a patient will be readmitted. The stay is between 1 and 14 days with a median of 4. This shows that a patient will regularly have to

Theo Jagodits 21416074
Grant Liu 93027462
Neal Sharma 38568337

stay to be monitored which indicates pretty serious incidents.

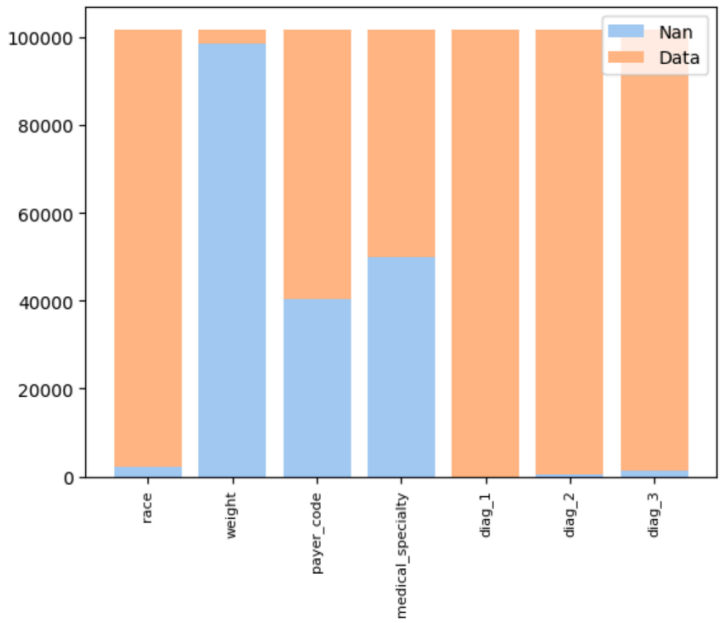
The number of medications is not the number of medications the patient is on, but the number that are administered during the patient stay. Typically an average of 16 medications is administered with a maximum value of 81 in the dataset. The most popular is by far insulin (see appendix). It is not mentioned in the data set summary if the medications are specifically for diabetes or other complications as well. We will assume that there might be other complications with patients for this report.

The number of diagnoses is another feature worth exploring. A minimum of one diagnosis is of course present since the patient would have diabetes. However the median is 8 which seems to allude that patients have a variety of complications, not just diabetes. This could also explain the high number of medications as mentioned earlier.

The last set of features is about the treatments given to the patients which range from popular diabetes medications. The top three medications for this feature subset are: insulin, metformin and glipizide. No further analysis is done here, these features consist of No, Up, Down, and Steady values corresponding to whether it has been given to the patient and whether the dosage has changed.

Feature Engineering

For feature selection, we split up the features based on whether they were categorical or numerical values. For each of the categorical features we used a Chi-Squared test on the target readmittance values to evaluate if there was a significant relationship. The Chi-Squared test is useful here because it does not depend on real values and makes no assumptions about the distribution of the data. We used an alpha of 0.05 which gives us a confidence level of 95% that the data is statistically significant. Running this test we manage to reduce the amount of categorical features from 40 to 25. All of the features that failed the test were diabetic treatments. A table will be included in the appendix.



The numerical treatments received the same process as with the categorical except a different statistical test was used. Here we used the Mann-Whitney U test which can be used with non-normally distributed data and a binary categorical variable to see if they are related. We plotted the numerical data to confirm that the data is not normally distributed (see appendix). After confirming, we also change the readmitted value from ‘<30’, ‘>30’ and NO to the variables ‘1’ and ‘0’ respectively to represent whether they have been readmitted. We then performed this test on all numerical features which all passed with 95% confidence. The results

are listed in the appendix for further inspection.

The next issue is missing values. After scanning the data set abstract we can conclude that only 7 features contain missing values. As we can see from the graph, weight is almost completely Nan and payer code and medical are both missing 40% and 50% missing respectively. We chose to completely remove weight since most of the values are missing as well as payer code and medical speciality. The other missing values are all less than 2% so we chose to keep those in.

Model Selection

We will be using two different machine learning techniques, Random Forest and Neural Networks. Both are very versatile techniques, and we believe that they will perform well on our data for the following reasons:

Random Forest is an ensemble learning method that combines the predictions of multiple individual models (decision trees) to improve overall performance. The ensemble nature helps in reducing overfitting and increasing the model's generalization ability. By aggregating the predictions of multiple trees, it can achieve a high level of accuracy despite using a “simple” algorithm. It can also capture complex relationships and patterns in the data, which is helpful for classification, and it doesn’t require the data to be scaled by the nature of decision trees. Random Forest is also robust to outliers and noisy data since the combination of predictions from multiple trees tends to smooth out the impact of individual outliers. Another bonus is that RF inherently performs feature selection since decision trees will split on the best features, although we have already put much effort into feature selection. For these reasons, we believe that Random Forest will perform well on our data

Neural networks can automatically learn hierarchical representations of data. Each layer in a deep neural network learns increasingly abstract features, allowing the model to capture complex patterns and relationships in the data. As a result, they do a good job of capturing non-linear relationships and are able to generalize well to new data. They can also filter out irrelevant features, which helps to deal with noise that is especially present in real-world data such as ours. Since neural networks are such a flexible technique, we believe that they may do a good job of learning the relationships in our data and classifying the correct labels. However, a possible challenge is choosing the right architecture to build our network with.

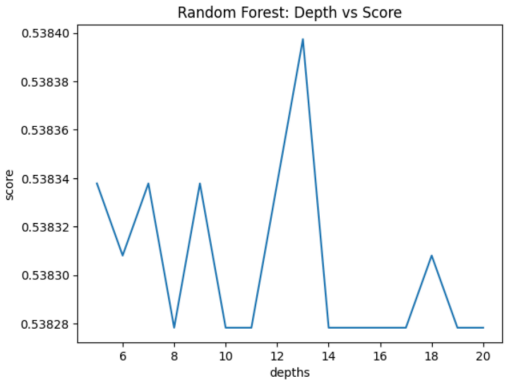
Model Tuning and Performance

The initialization of parameters for machine learning models is a large part of how well a model performs on a given set of data. We decided to check different initial parameters for initializing our models. We used an 80-20 training-testing split on our data after one hot encoding all of the necessary statistically significant categorical features that we established in the previous section.

Group 27

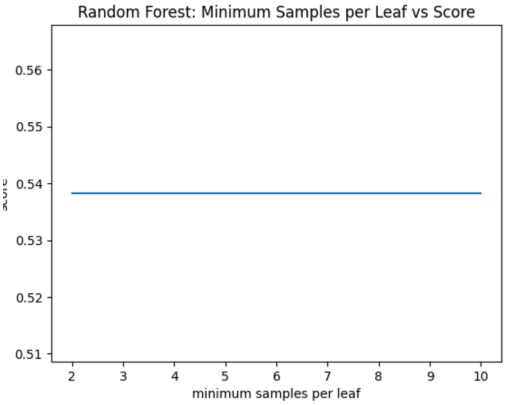
Theo Jagodits 21416074
Grant Liu 93027462
Neal Sharma 38568337

Random Forest



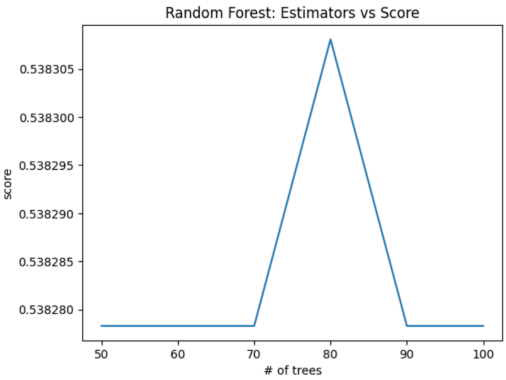
We tried a range of different tree depths in our ensemble to see what number would perform the best. However, given that the score changes whenever the block is rerun, and the score barely varies between the given range of estimators the default value of depth 10 seems to be serviceable. In this case a depth of 12-13 seems best, but this is not always the case.

We also tried to vary the number of trees in our ensemble. Given the range, it seems 80 trees in the ensemble returns the best result. Varying the minimum required number of samples per leaf seems to have no effect on increasing the accuracy of our model.



Using the overall best performing initial parameters as suggested by the model tuning above, it seems that we can reach an accuracy of 53.83%. While there is certainly more room for improvement, if we had a more powerful machine perhaps we could test more parameters to fine tune our model even better.

Neural Network



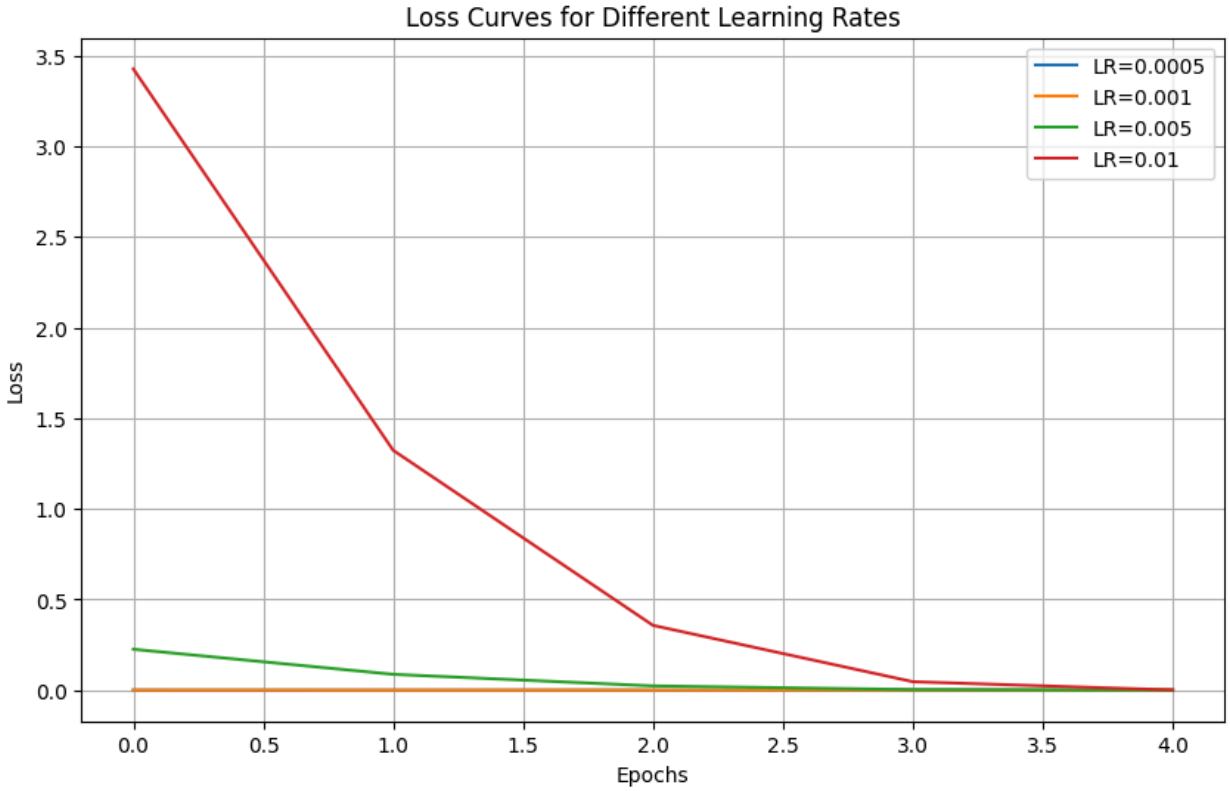
For the neural networks, we wanted to test the effect of different learning rates on determining the readmittance rate of patients. We first standardized the features using the StandardScaler(). The learning rates used were [0.0005, 0.001, 0.005, 0.01].

The neural network model was parameterized with a sequential convolutional neural network with four dense layers, the first three having 256, 128, and 64 nodes and ReLu activation functions. The output layer consisted of a dense layer with one node and a sigmoid activation function. Using the Adam optimizer, the model was compiled with cross entropy as the loss function. We then trained and fit the model for 5 epochs on a batch size of 256 and used the validation data to produce the error curves.

Initially, the accuracy was 33% when we had the batch size set to 1024, the model was overfitting and was not converging fast enough. Additionally, because of the low learning rates, smaller batch sizes were necessary to avoid diverging during training. Smaller batch sizes in neural network training introduce more noise in gradient updates. To maintain stable training, smaller learning rates are often used to avoid divergence caused by this noise. This trade-off helps achieve a balance between faster convergence and stability while guarding against overfitting.

Theo Jagodits 21416074
Grant Liu 93027462
Neal Sharma 38568337

After changing the batch size to 256, we were able to get an average accuracy of 81%. Additionally, the 0.01 learning rate had the highest drop in loss, but the lower learning rates had much lower initial loss patterns.



Future Work

Considering future work for the neural network model, if we had a more powerful server, we would run more epochs per trial. In our study using the UCI dataset, we encountered certain resource limitations that present opportunities for future research. Given our restricted access to computational resources, we were constrained in the complexity of our machine learning models and the scale of hyperparameter tuning. In future work, securing access to high-performance computing clusters or cloud resources could enable us to explore more sophisticated model architectures and conduct extensive hyperparameter optimization, potentially leading to further performance improvements. Additionally, due to the inherent class imbalance in the dataset, we employed basic oversampling techniques to address this issue.

Future research could delve deeper into advanced techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or anomaly detection methods to better handle imbalanced data. Furthermore, we acknowledge that the UCI dataset captures a limited time frame, and longitudinal data collection or integration with other related datasets could provide a more comprehensive understanding of the underlying patterns and trends in the domain. Finally, as ethical considerations are paramount, future work should involve a thorough examination of

potential biases in the dataset and the development of fairness-aware models to mitigate any disparities in predictions.

Future research for our Random Forest classifier could involve training different parameters to see how much they affect our model. Unfortunately, there was a large problem with high run times and lack of RAM, so we had to simplify our model tuning to the process that we've used above. In the future it would also be nice to have access to more powerful machines to be able to test more features.

Appendix

We opted out of an appendix and will be submitting a zip file with our additional graphs and tables.

Work:
Theodore did Feature Analysis and Feature Engineering
Grant was in charge of Random Forest
Neal did Neural Net
We collaborated on all aspects of planning and then divided work.