Slide 1:

Aoa! Im here to present my last month work for Type 2 Diabetes prediction model

Slide2: Overview

First there is a brief overview of the presentation that what I’m covering today.

First there will be quick recap of work done before July then the challenges that we have faced during the working month. After that there is some changes that we have made in our dataset using the provider notes. There is a detailed overview of the feature selection process. And some possible future directions for the prediction model.

Slide 3: Patient Distribution

So, we have a dataset of around 36000 patients including 5000 Diabetic and 31000 non-type 2 diabetic patients. The patients we have are shortlisted based on if they have at least 3 encounters.

Slide 4: Classifiers Evaluation

There are some initial model evaluations that we have done on our dataset that was comprising of patient’s demographics, vitals and diagnosis. And we have shortlisted these top performing models on our dataset which includes 2 linear and 2 ensemble methods

Slide 5: Feature Scaling

Then we have modified the feature scaling step in our models according to the requirement of the model for example the distance-based models like SVM and KNN require standardization in which the distribution have 0 mean and 1 standard deviation.

Slide 6: Challenges

Slide 7: Mice imputation

There is a presence of missing values in the vitals so it was previously resolved by imputation with exponential weighted average but there is a more promising approach that I have recently explored is the MICE imputation which predicts missing values using known values from other features.

Slide 8: Mice Imputation

What it does is it first imputes the missing values by the mean imputation and takes each feature one by one removes the mean imputed value and then predicts that value by the known values until all the features mean imputed values are replaced by the predicted values.

The algorithm keeps on repeating for the fixed number of iterations. There are two variants one is with iterative imputation and the other is LightGBM method that I haven’t explored yet.

Slide 10: Provider Notes

Slide 11: Medications

We have shortlisted some of the Medications from literature that increases the risk of type 2 diabetes causing side effects on glucose metabolism that how the glucose is converted to energy, causes insulin resistance or causes weight gain.

Slide 12: Procedures

Similarly, we have listed some surgeries as well that have effect on the insulin production or have any negative effect on the glucose metabolism

Slide 13: Outliers

After extracting all these features, we also have to handle the outliers as after extraction the distribution also changes accordingly so the weight and height that lie outside this range are considered outliers and we remove those patients from our dataset.

Slide 14: Results

Curse of dimensionality

Slide 15: Boruta

After this we initiated the feature selection process with Boruta feature selection technique.

* Shuffled copies of the features
* Train random forest on the extended data and applies feature importance measure that is mean decrease in accuracy.
* Features that have significantly higher importance in the original dataset compared to the shadow dataset are considered relevant and are retained, while others are eliminated.

Slide 16: Mutual Information

* Mutual Information Gain is based on the concept of information entropy from information theory. Entropy represents the amount of uncertainty or randomness in a variable. If a variable has low entropy, it means its values are more predictable and provide valuable information.
* To calculate the mutual information between two variables, you compare their joint distribution (how often their values occur together) with what you would expect if the variables were independent (not related). If the joint distribution is similar to the expected distribution, the mutual information is low. If it's different, the mutual information is high.
* In the context of feature selection, you calculate the mutual information between each feature and the target variable you want to predict. Features with higher mutual information are considered more important because they provide more information about the target variable.

Slide 17: Chi2

* The chi-squared test is a statistical tool used to determine if there's a significant relationship between two categorical variables. It helps us understand if the differences in observed data are likely due to a real association or just random chance.
* The test involves comparing the expected frequencies of observations in different categories with the actual observed frequencies. It calculates a chi-squared value, which is then compared to a critical value from a chi-squared distribution. If the calculated value is significantly larger than the critical value, it suggests that the variables are not independent and likely have a meaningful connection.
* like "favorite ice cream flavors" and "gender."

Slide 18: Results comparison after feature selection

Reduction in Noise

Slide 19: Summary

Slide 20: Future Directions

Slide 21: Thank you!