Predicting Hemoglobin A1c Levels in Diabetes Patients Using Frequency of Visits

Danielle Mizrachi

Problem Statement

A non profit clinic wants to know whether or not frequency of patient visits has an impact on their overall A1c level.

Questions

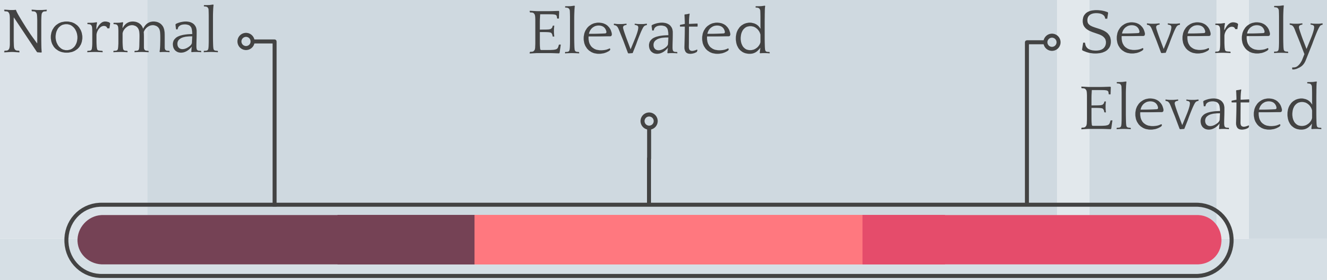
Are patients more likely to have a lower and controlled A1c if they visit more frequently?

Can we predict based off of frequency of visits or time in between visits, what the patient's A1c is?

Might factors like age also have an impact on A1c levels?

A1c Levels

“The test results give you a picture of your average blood sugar level over the past two to three months. The higher the levels, the greater your risk of developing diabetes complications.” – ADA



Normal

Elevated

Severely Elevated

4-6

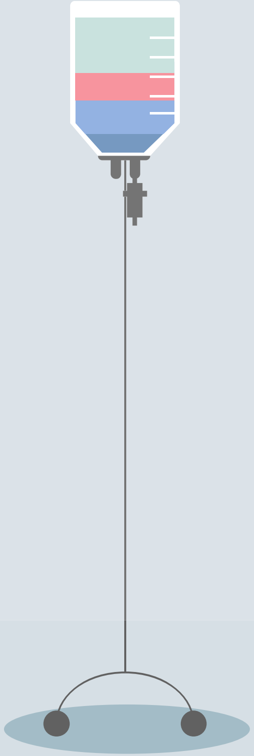
7-9

10-13

Controlled

Uncontrolled

The Process



01

CLEANING THE DATA

02

SPLITTING THE DATA

03

FEATURE ENGINEERING

04

EDA

05

MODELING

06

NLP

Data Cleaning and Splitting

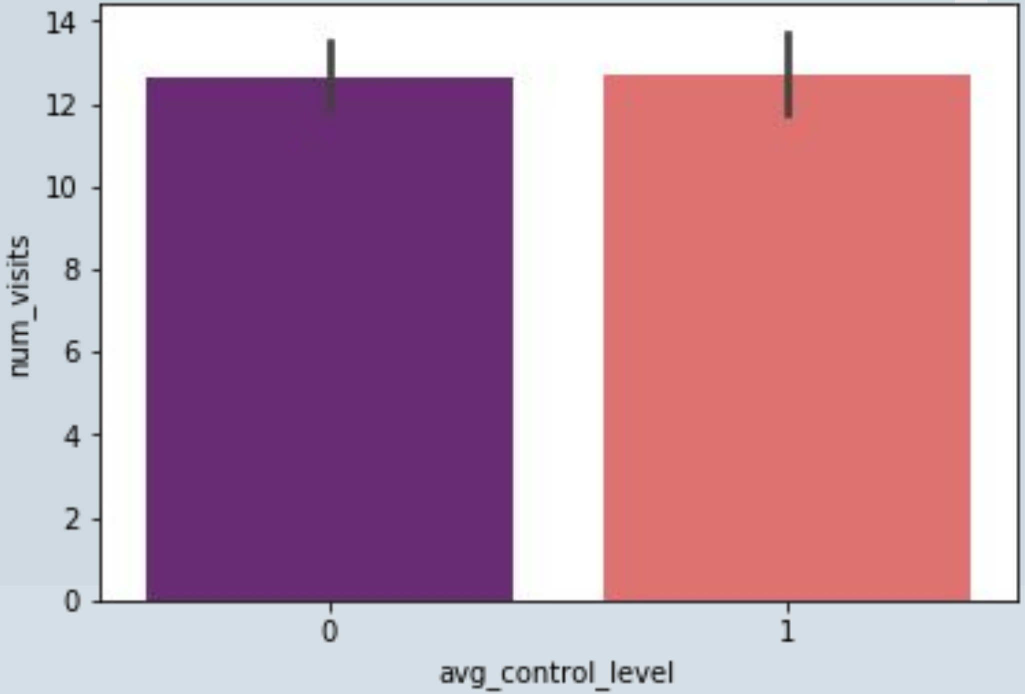
- We started with 273,485 rows of data and 8 columns. This data represents 1,099 patients
- Dropped rows where A1c is null, since there were only 32
- Changed data types to represent datetime columns (observation date and visit date) and objects to floats.
- Removed or fix outliers: found a minimum too low for A1c. Changed .089 to 8.9 and dropped 1.7 a1c levels.
- Created an additional column that marks each patients A1c as controlled as our positive class 1 or uncontrolled as our negative class 0

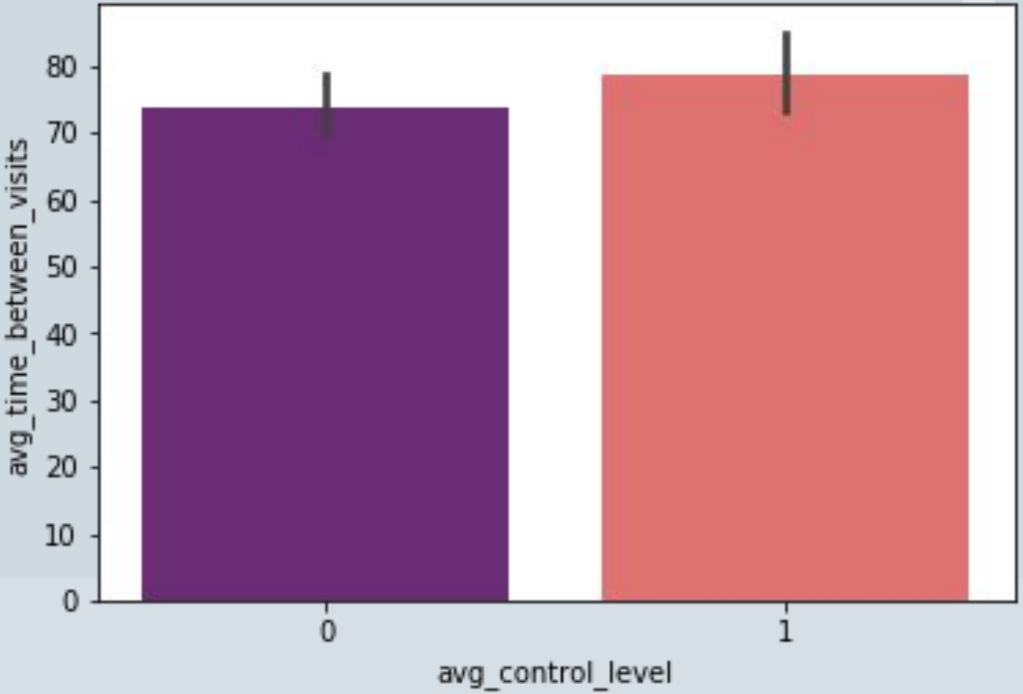
Data Splitting

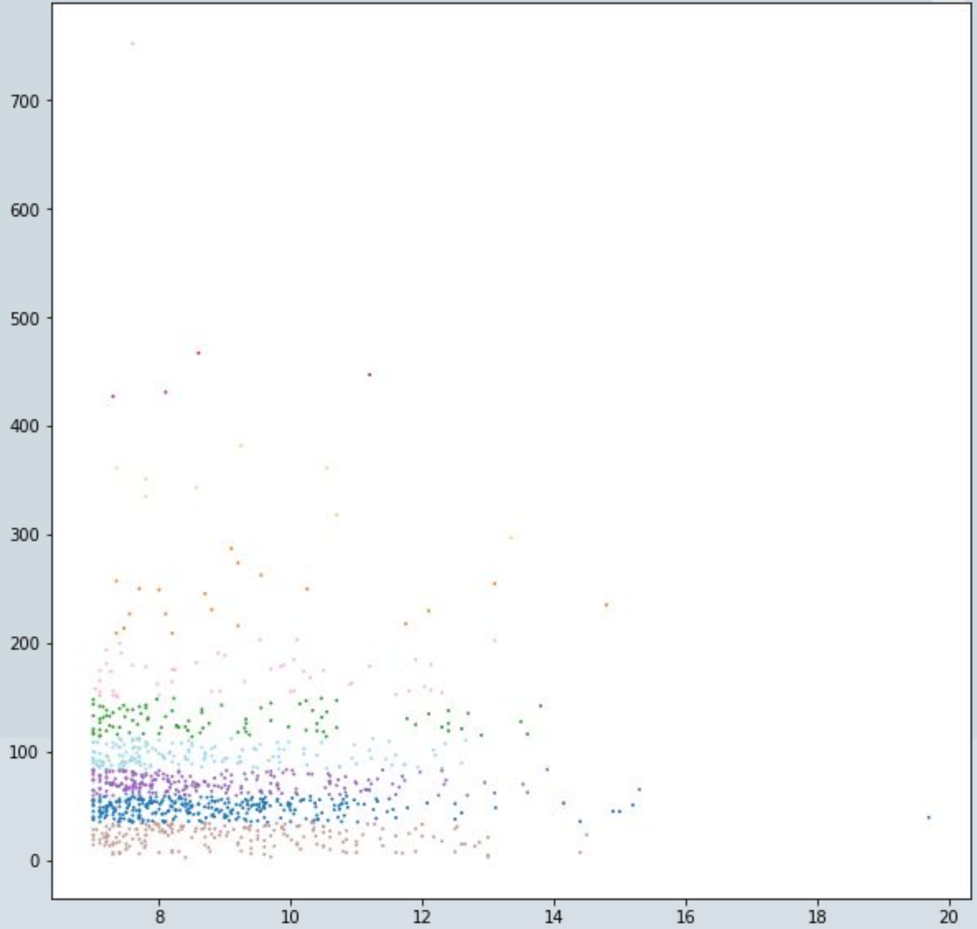
- This data was not normalized, there are two dataframes grouped together in a way that makes it difficult to analyze.
- One is responsible for a1c and when blood work was done, and the other holds information only on office visits.
- This also created many duplicates in the original dataset - I split the data into two dataframes and dropped the duplicates in each.
- In doing this, I discovered that people come in a lot more frequently than they get their blood work taken.

Feature Engineering

- I created a function to count how many unique visits and unique observation dates each patient had
- I built a function that calculated time between visits, and then from this the average time in between visits to be used as a feature
- To model using NLP, I combined and stemmed all summary text together and added it to each respective patient's row.
- I built a function to add average A1c of each patient as an additional column, and from this added a column to represent a patient's average control level.


14
12
10
8
6
4
2
0
num_visits
0
1
avg_control_level


avg_time_between_visits
0
10
20
30
40
50
60
70
80
0
1
avg_control_level

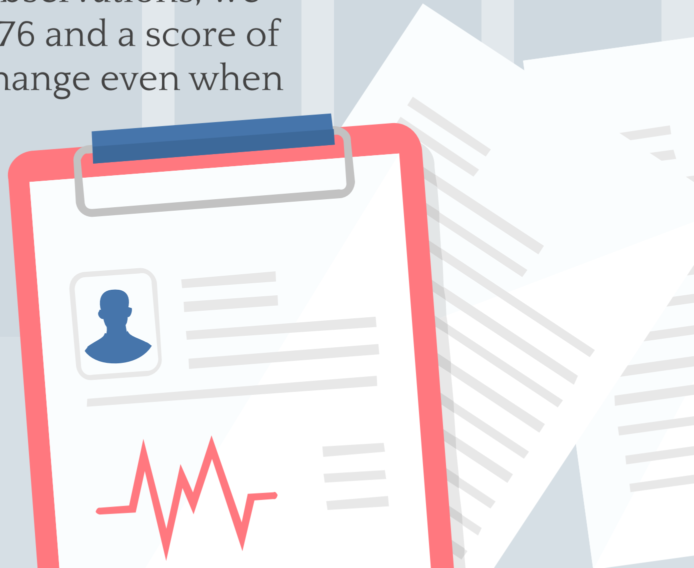

700
600
500
400
300
200
100
0
8
10
12
14
16
18
20

Modelling:
Regression

Linear Regression

How well does the model predict average a1c if we just use average time between visits and number of visits for our features?

We get a score on the training set of 0.006 and 0.0116 on the testing set. If we add in age and number of observations, we still get a low score on the training set of 0.076 and a score of 0.088 on the testing set. The score doesn't change even when we round our a1c levels.



Modeling: Classification

Logistic Regression

Score on the training set: 0.67
Score on the test set: 0.63

Decision Tree Classifier

Score on training set: 0.66
Score on testing set: 0.57

Sensitivity: 0.3652
Specificity: 0.7237

Random Forest Classifier

Score on testing set: 0.645
Sensitivity: .206

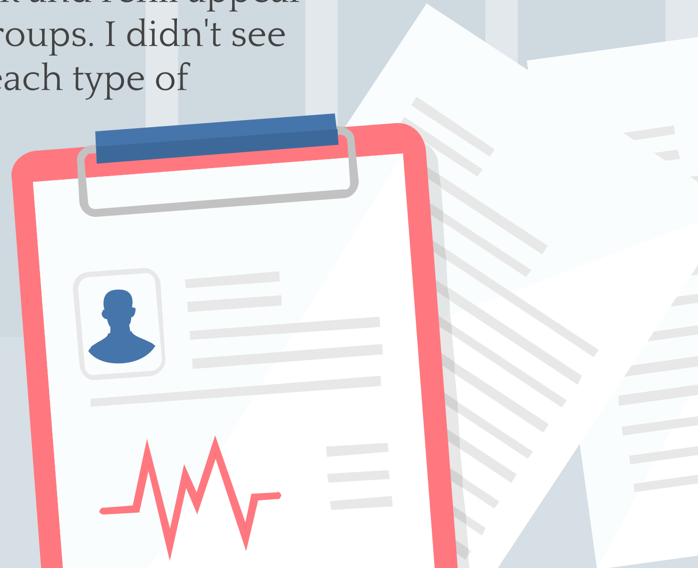
Modelling: NLP

Natural Language Processing

How well does the model predict average a1c if we just use the sum of the summary text of each patient's visits?

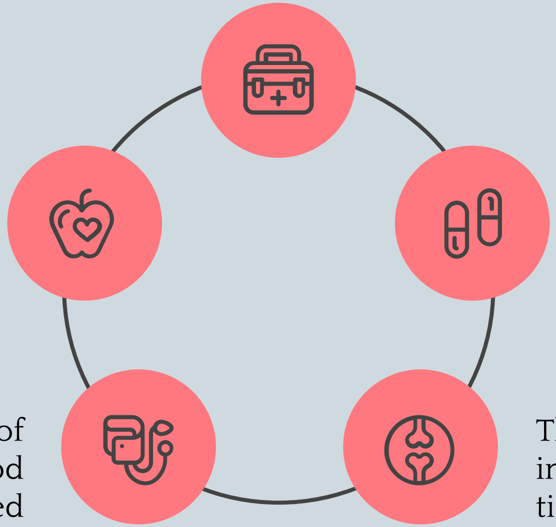
Our NLP best score was .635.

Words like hypertension, diabetes, lab, check and refill appear to be in both controlled and uncontrolled groups. I didn't see any major differences in language used for each type of patient.



Conclusions and Limitations

It is difficult to make any solid conclusions from this analysis given the limitations of the project and lack of knowledge on the subject matter.



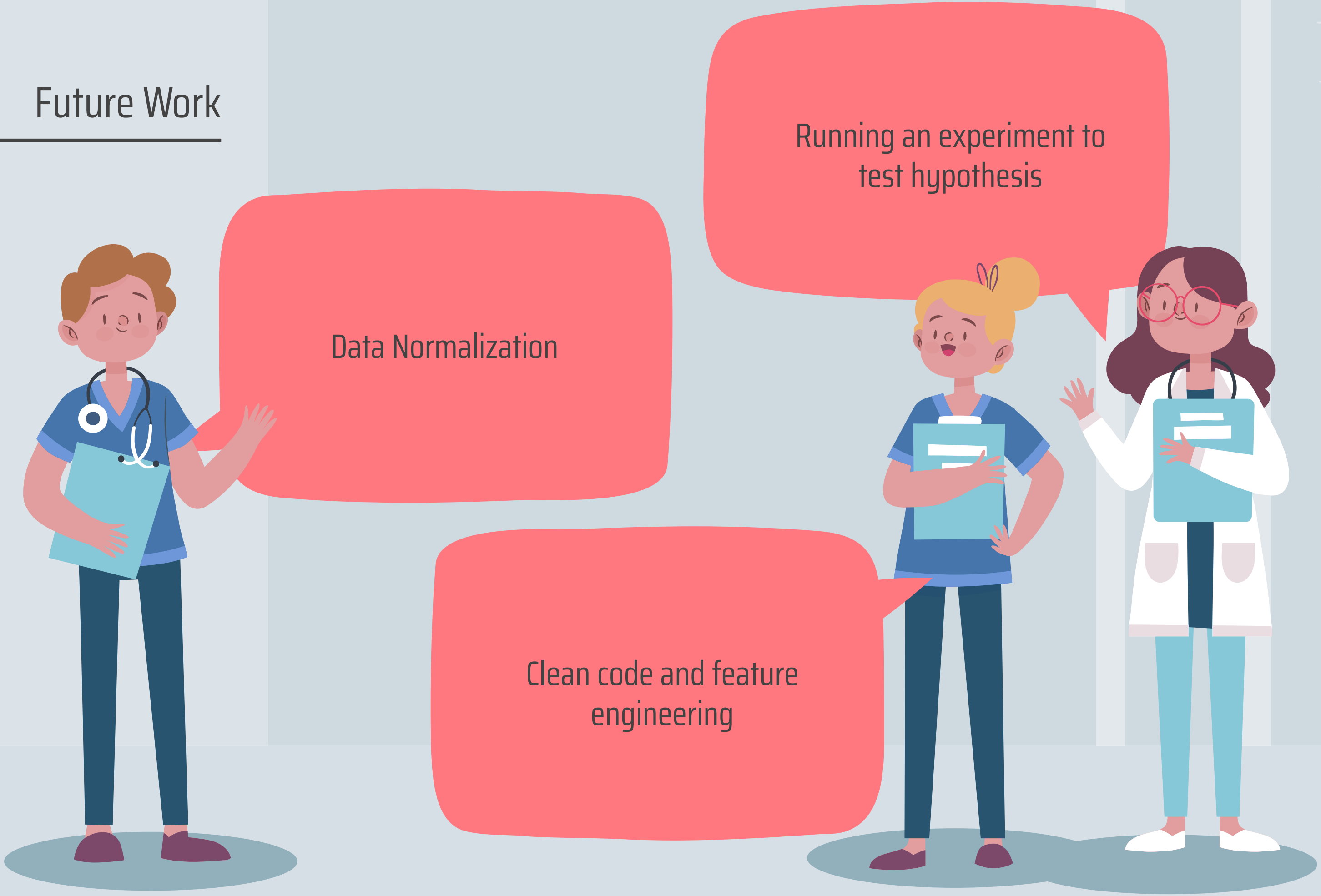
Having to manipulate the data also doesn't seem like the optimal approach given how unreliable averages can be at times.

There was perhaps some indication that more time in between visits was correlated with lower average a1c levels, but we'd need to investigate this further in a controlled experiment.

The inconsistency of visits and when blood work drawn also acted as a limitation.

The task at hand was more challenging given how the data was gathered, stored, and pulled.

Future Work



Data Normalization

Running an experiment to test hypothesis

Clean code and feature engineering

Questions?

THANKS

CREDITS

- Presentation template by Slidesgo
- Icons by Flaticon
- Infographics by Freepik
- Images created by Freepik
- Text & Image slide photo created by Freepik.com