Datathon: Diabetes Treatment

Team 18

April 2022

1 Executive Summary

Diabetes is a widespread disease in the US and results in significant expenditures for the healthcare system. Furthermore, due to the segregated nature of the healthcare system in the US, different patients are covered by different types of insurance plans or may not be covered at all. A major point of debate in the country is the effectiveness of the healthcare system and whether standards of care vary across coverage types. Thus our team has decided to focus our research on answering the following question:

How does insurance type influence the speed and effectiveness of diabetes treatment ?

As 30 percent of the aggregate expenditure is dedicated to in-patient care, our investigation can reveal how the insurance type affects the time patients spend in in-patient care and how often patients are readmitted. Having an understanding of these factors can help identify where deficiencies in the healthcare system lie and potentially lead to development of more efficiencies that decrease overall expenditures on diabetes.

Our study of the data concludes that having private insurance leads to lower readmission rates compared to having public insurance/self-pay. We also found that there is a positive but weak relationship between having public insurance/self-pay, and the time spent in hospital. These findings prompt larger questions about the effectiveness of public health insurance in the US and further research could be conducted to identify the micro factors that could be improved upon. The result of additional research in this field could be better health outcomes for diabetic patients and reduced public healthcare expenditures.

2 Technical exposition

2.1 Data cleaning and feature engineering

2.1.1 Initial cleaning

We chose to focus on unique patients (as to satisfy i.i.d assumptions of regression modelling), therefore reducing the dataset from 100k records to approximately 70k by dropping patient_nbr duplicates. Thereafter, in a similar fashion to (Strack et al., 2014), we removed patients who end up dying or discharged to hospice as they would not count as re-addmited even though they have the worse health outcome. Furthermore, we chose to discard rows which contained erroneous/missing values in term of race, gender and age.

2.1.2 Encoding Insurance Codes

Our first approach to solving the question was analysing the payer code column. The dataset initially contained 18 different types of payer codes. We mapped them into three categories in order to reduce imbalance and bias : Public Insurance, Private Insurance and Self-Paid. We also dropped missing values (around 40k rows). Here you can visualize the distribution of the 3 categories.
In order to turn this into numeric data, we tagged Public and Self Pay types as 0 and Private as 1. This is as Public/Self Pay generally results in worse coverage ((Self-Funded, Non-Federal Governmental Plans — CMS, 2022).

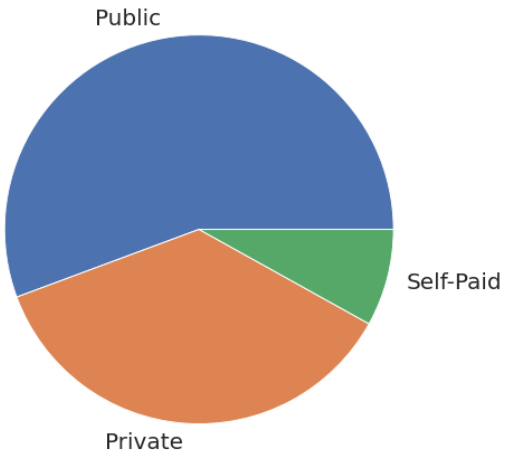

Figure 1: Distribution of the new categories

2.1.3 Hospital Re-admissions

When analysing the quality of healthcare, the readmission rate is one of the most important indicators that governmental agencies watch out for. As pointed out in a ScholarWorks article, measures have been taken by governmental agencies to reduce the 30 day readmission rate. In fact, hospitals with high excess readmission ratios (ERRs) are subject to a financial penalty which results in healthcare providers having a strong incentive to reduce the ERR.
As it is an important variable in determining the effectiveness of treatment, we decided to choose this as our target variable.
In our dataset, the readmission column was split into 3 categories: "not readmitted", "readmitted in under 30 days", and "readmitted after 30 days". For our modelling process, We decided to bucket this further into 2 categories, "readmitted" and "not readmitted" to allow us to carry out a binary classification.

2.1.4 Age groups

The age groups in our dataset was split into 10 categories, which we further converted into numerical data by taking the midpoint of each given interval.

2.1.5 Admission type

The admission type id specifies the conditions under which each patient was admitted to the hospital. For example, we have the following categories: "elective", "urgent", "emergency", "Newborn" etc.. About 38 000 patients fall into the following categories: "Emergency", "Urgent", "Elective", so we decided to map these values to an ordinal scale as follows:

- 2 : Emergency
- 1 : Urgent
- 0 : Elective

This feature is particularly important as it serves as an indicator concerning the severity of diabetes of each patient upon hospital admission. Values which did not fall into these categories were taken to be the mean ordinal value of 1.
We are going to approach our hypotheses through the lens of two models : a Logistic Regression using a subset of inputs to interpret readmission (**effectiveness**) and a Linear Regression using a similar subset to interpret time spent in the hospital (**speed**).

2.1.6 Feature Engineering for Continuous Variables

We chose to standardize our continuous data as:

$$Z = \frac{x - \mu}{\sigma}$$

which results in independent variables that have a mean of 0 and a standard deviation of 1. This step is important for our modelling process as it ensures that all of the input variables to the model are on the same scale. This step has further benefits for modelling by leading to regression coefficients that are simpler to compare against each other (SIEGEL, 2022).

2.1.7 Feature Engineering for Categorical Variables

We used one-hot encoding in order to stabilise and fit our categorical variables to our model (Brownlee, 2022). This was applied to variables such as race or diagnostic types in order make our models capture the global context of the observations.

2.2 Preliminary insights on the relationship between insurance and effectiveness of treatment

2.2.1 Are patients who have public insurance more likely to be readmitted?

The following bar plot shows that patients who have public insurance are more likely to be readmitted in hospital. More precisely, out of the privately insured patients, 34% were readmitted, whereas for publicly patients people, 43% were readmitted.

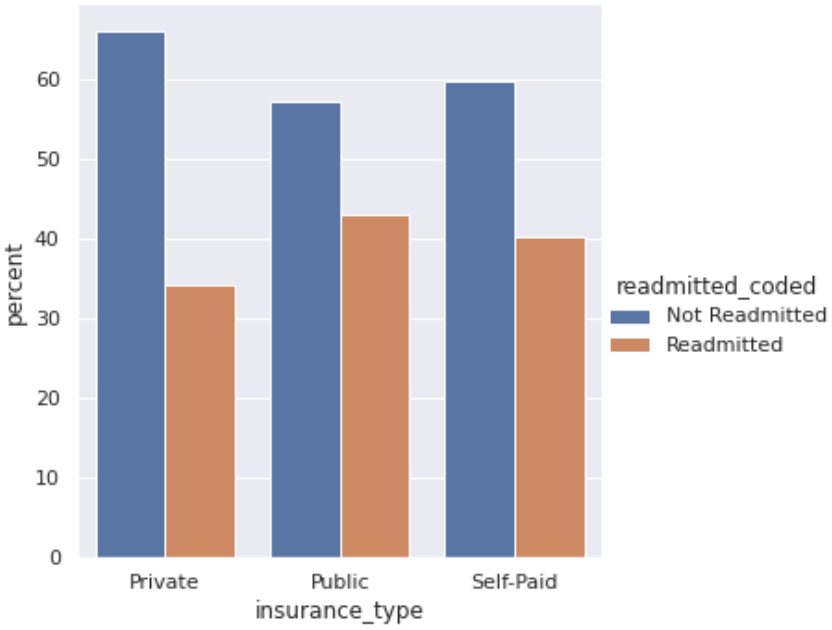

Figure 2: Readmission by type of insurance

2.2.2 Effect of Insurance Plan on Drugs Administered

The first hypothesis we made is that the discrepancy in effectiveness of treatment is due to a difference in the medication that is given to patients according to their insurance type. In other words, are patients with public insurance given different drugs than patients who are privately insured?

The following correlation matrix shows that the type of drugs across all categories is similar, so we can safely assume that the type of insurance of the patient does NOT influence what type of drugs they receive.

For this reason, we decided to not include the types of drugs given to the patient as part of our regressors. This is as we believe the prescription of drugs between insurance is homogeneous and would add unnecessary additional complexity to the model

	Private	Public	Self-Paid
Private	1	99.1%	99.6%
Public	99.1%	1	99.3%
Self-Paid	99.6%	99.3%	1

Table 1: Correlations of drug prescription per Insurance Type

2.2.3 Demographic Implications - Age

In order to correct the bias that stems from socio-demographic factors, we have decided to include variables such as age, race and gender in our regression. For example, when looking at the distribution of age in each category, we can clearly see that publicly insured patients are usually older than in other insurance categories.

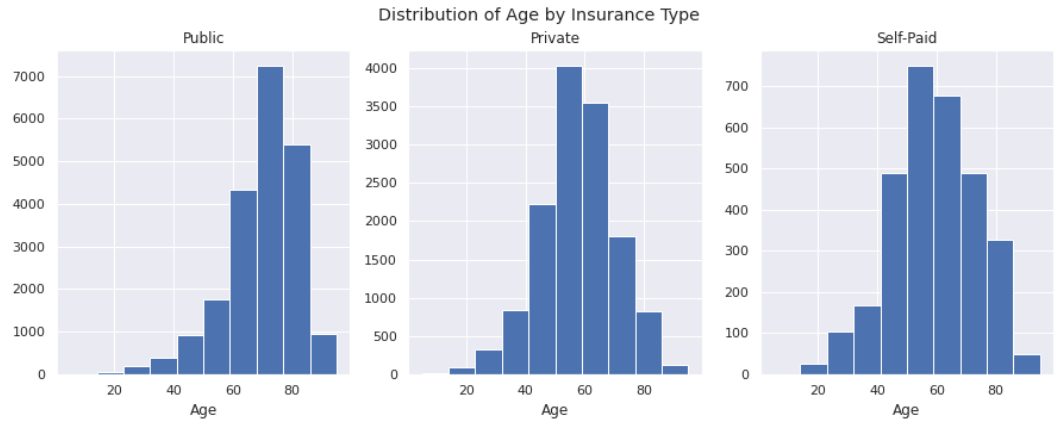

Figure 3: Distribution of age groups across insurance categories

2.2.4 Demographic Implications - Gender

In the figure of the gender distribution across insurance types, we see that for private insurance there is an even split between male and females. For public insurance however, there is a clear imbalance between the gender splits with a greater proportion of people on public insurance being female vs. male. For those with self-paid insurance however, there is a higher proportion of males vs. females.

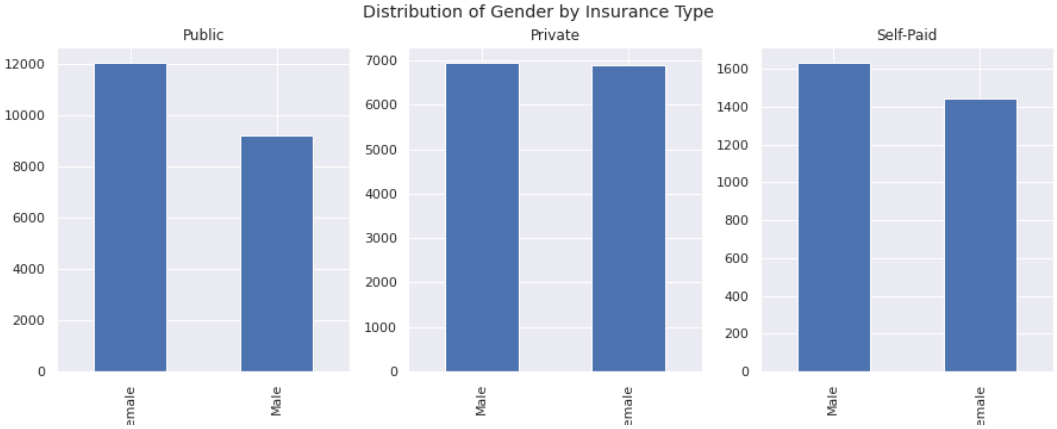

Figure 4: Distribution of gender across insurance categories

2.2.5 Demographic Implications - Race

In the figure of the race distribution across insurance types, we do not discern any major differences. The distributions are fairly consistent across the 3 insurance types so we do not expect race to be an important factor in our model.

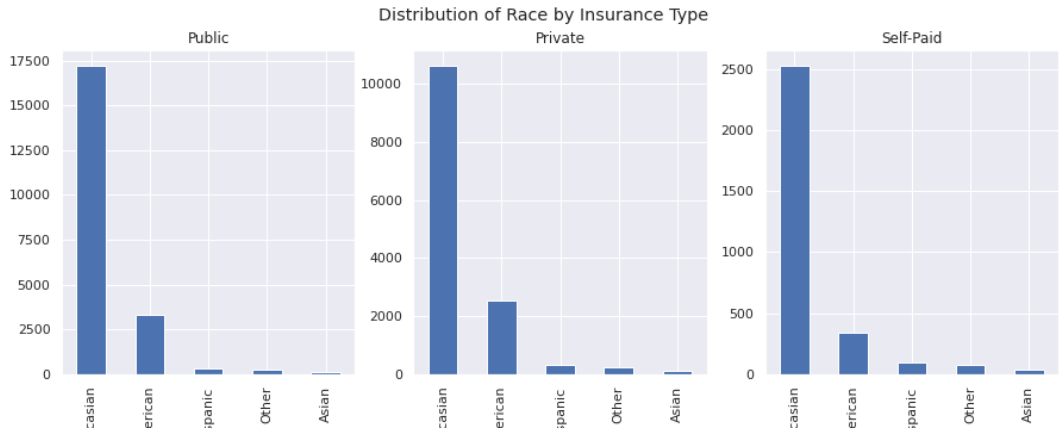

Figure 5: Distribution of race across insurance categories

2.3 First insights on the relationship between Time spent in Hospital and Insurance Type

The amount of time spent by patients in the hospital is an analogue for the speed of treatment.

We infered from the cleaned dataset that there was a marked difference in the time spent in hospital between each insurance type. The average number of days spent in hospital is lowest for private insurance at 3.8 days and highest for public insurance at 4.4 days. Self-Paid insurance is similar to Private insurance at 3.9 days. This suggests that the speed of the treatment could be faster for private and self paid insurance than public insurance. We hypothesize that this could be due to private insurance being more efficient than public insurance and that the care provided privately results in quicker improvement in diabetic conditions. We therefore chose to avoid confounding effects and included time spent in the hospital as a predictor for readmission (**effectiveness section**).

We also chose to create a regression in order to estimate the influence of insurance type on time spent in the hospital (**speed section**).

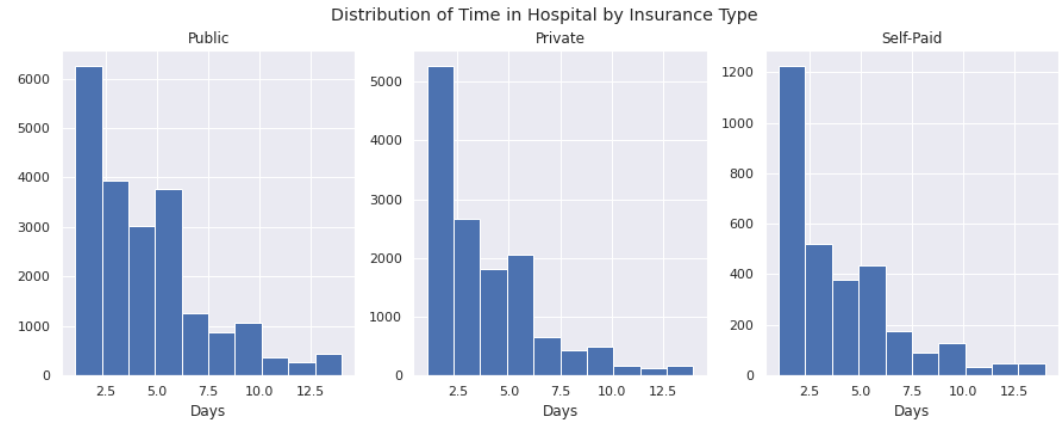

Figure 6: Distribution of time spend in hospital across insurance categories

	Mean	Standard Deviation
Private	3.8	2.7
Public	4.4	2.9
Self-Paid	3.9	2.8

Table 2: Time in Hospital by Insurance Type

3 Feature Choice and Modelling

3.1 Dropped data

As the following features did not seem relevant in our study or were too sparse, we decided to drop the following columns: admission source and medical specialty of the doctor. Additionally, as explained in the previous section, the types of drugs administered were homogeneous across insurance groups, therefore could be dropped. We also had to drop the variable of the weight of the patient, as the majority of data was missing for this variable.

3.2 Effectiveness Hypothesis

3.2.1 Variables

Based on our explanatory data discovery and hypotheses, we decided to test for "effectiveness" related to insurance type through the following features :

Input	Dtype
age	float64
admission_severity	float64
time_in_hospital	float64
num_lab_procedures	float64
num_procedures	float64
num_medications	float64
number_outpatient	float64
number_emergency	float64
number_inpatient	float64
number_diagnoses	float64
insurance_type	int64
race_Asian	uint8
race_Caucasian	uint8
race_Hispanic	uint8
race_Other	uint8
diag_type_Diabetes	uint8
diag_type_Digestive	uint8
diag_type_Genitourinary	uint8
diag_type_Injury	uint8
diag_type_Musculoskeletal	uint8
diag_type_Neoplasms	uint8
diag_type_Other	uint8
diag_type_Respiratory	uint8
gender_Male	uint8
change_No	uint8
diabetesMed_Yes	uint8

Table 3: Input and Input Types

We therefore infer severity at admission time through 'admission_severity' as max_glu_serum or A1C results, had a large majority of missing data.

As previously mentioned, we decided to use a binary interpretation of readmission in the form of

$$0 = \text{no readmission}, 1 = \text{readmission before or after 30 days}$$

3.2.2 Model Choice, Results and Statistical Insights

Our target variable for readmission was a discrete binary response. Since our observations were essentially i.i.d (dropping duplicates) and that we had low colinearity between factors (maximum was 47% in our final dataset) with a large sample size, we chose Logistic Regression in order estimate significance and importance of insurance type for readmission outcome.

There was found to be a negative relationship between private-insurance and hospital readmission rates. A Z-Value of **-11.253** was obtained for the relationship. This leads to a p-value of **2.23e-29**. This clearly indicates that there is a statistically significant relationship between insurance type and re-admittance rate.
In order to get an idea about the strength of the effect of having private insurance, the Beta of the

insurance variable needs to be compared to the Beta of other variables. However the regressors are a mix of numerical and categorical (binary) data types, and it is difficult to directly compare the Betas of these two types of data. Therefore the Beta of the Insurance Variable is compared to the Betas of the other categorical variables. The results are shown in the Bar Chart:

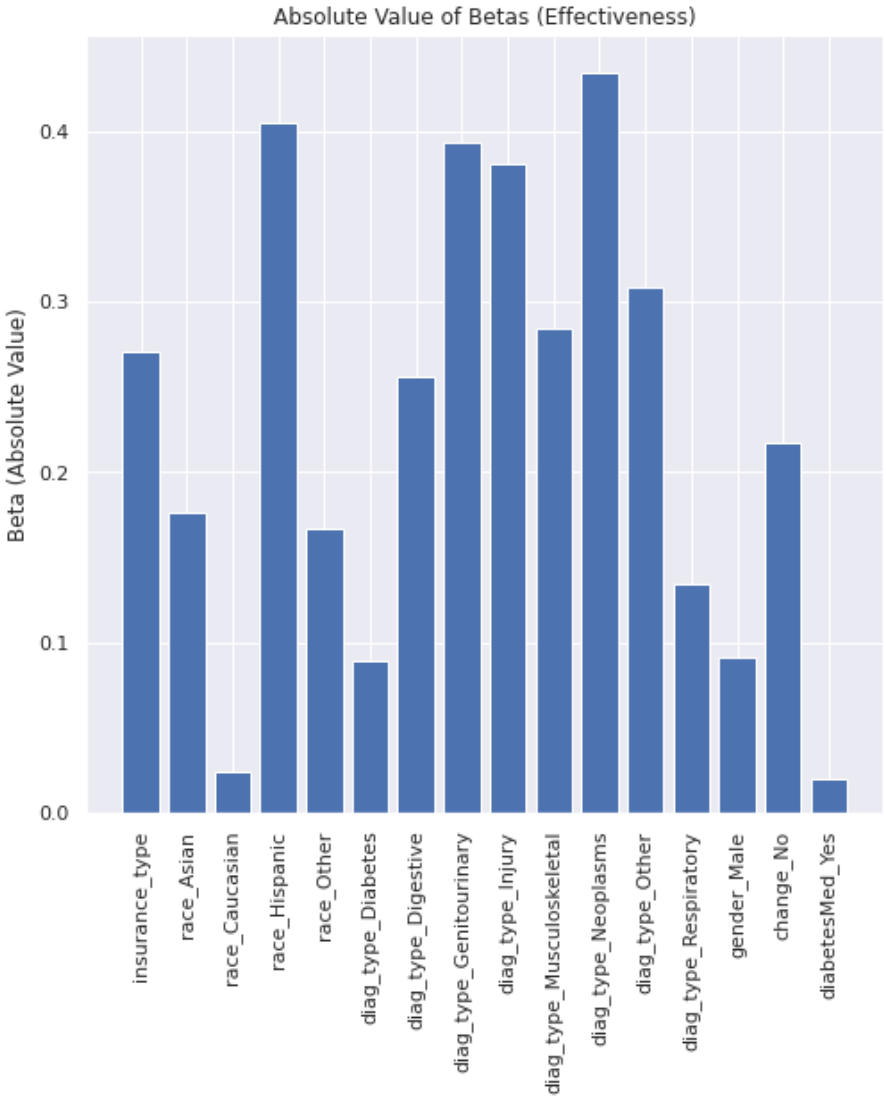

Figure 7: Absolute Value of Betas of Categorical Variables (Effectiveness Hypothesis)

We compared the absolute values of the Betas of different variables. The insurance variable is the first bar. We can see that the effect is fairly significant when compared to other variables. Some variables such as the diagnosis type have a greater effect, but we can see that the insurance variable **does not** have an insignificant effect.

3.3 Speed Hypothesis

3.3.1 Variables

The choice of variables for the linear regression is similar to that of Logistic Regression, we however swap **time_in_hospital** as our target variable and plug **readmit** (our Logistic target) back into our regressors. As we noticed in our EDA, there are demographic discrepancies between different insurance clients (Age mainly)

Input	Dtype
age	float64
admission_severity	float64
num_lab_procedures	float64
num_procedures	float64
num_medications	float64
number_outpatient	float64
number_emergency	float64
number_inpatient	float64
number_diagnoses	float64
insurance_type	int64
race_Asian	uint8
race_Caucasian	uint8
race_Hispanic	uint8
race_Other	uint8
diag_type_Diabetes	uint8
diag_type_Digestive	uint8
diag_type_Genitourinary	uint8
diag_type_Injury	uint8
diag_type_Musculoskeletal	uint8
diag_type_Neoplasms	uint8
diag_type_Other	uint8
diag_type_Respiratory	uint8
gender_Male	uint8
change_No	uint8
diabetesMed_Yes	uint8
readmit	int64

Table 4: Inputs for the Linear Regression

3.3.2 Model Results and Statistical Insights

Our target variable (time spent in hospital) was numerical and continuous. Our observations are i.i.d because we dropped duplicates. Our regressors are (as mentionned above) not strongly correlated. We also expect linearity to exist and choose to implement a **Linear Regression** model.

There was found to be a negative relationship between having private insurance, and the length of time spent in hospital. A t-Value of **-6.305** was obtained. This leads to a p-value of **-2.91e-10**. This indicates there is a statistically significant relationship between having private insurance and the length of stay at the hospital.
The R^2 of the model was found to be 0.316, this indicates that is is difficult to make direct predictions on the length of hospital stay. However we can still evaluate the t-values and Betas of the variables in a comparative analysis.
In the same method as before, a bar chart of the absolute values of beta for the categorical variables was plotted, and is shown below:

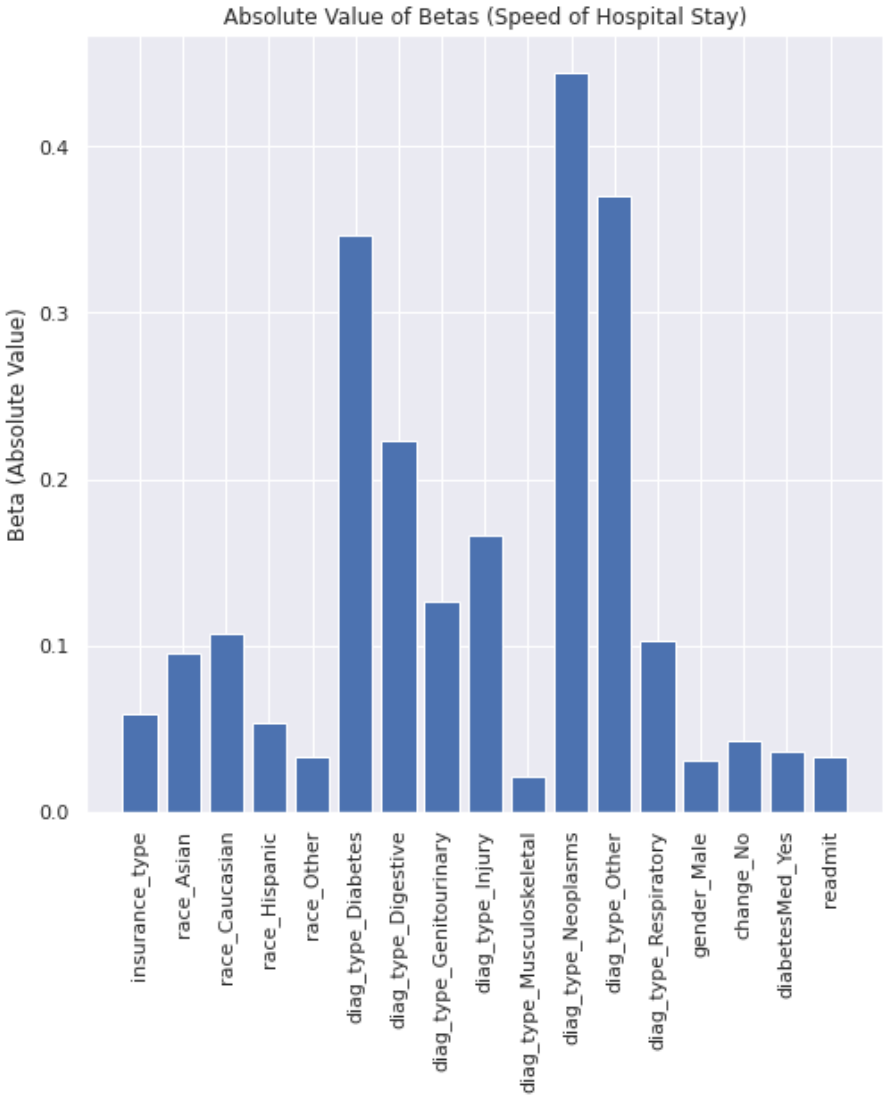

Figure 8: Absolute Value of Betas of Categorical Variables (Speed Hypothesis)

Unlike for hospital re-admittance rate, we can see that the strength of the effect of insurance type on time spent in hospital is small. The largest effects by far are to do with the diagnosis type for the patient.

A Bibliography

- Strack, B., DeShazo, J., Gennings, C., Olmo, J., Ventura, S., Cios, K. and Clore, J., 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, pp.1-11.

- Cms.gov. 2022. ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles — CMS. [online] Available at: ¡https://www.cms.gov/Medicare/Coding¿ [Accessed 9 April 2022]. pp.742-747.

- SIEGEL, A., 2022. PRACTICAL BUSINESS STATISTICS. [S.l.]: ELSEVIER ACADEMIC PRESS.

- Brownlee, J., 2022. Ordinal and One-Hot Encodings for Categorical Data. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ [Accessed 9 April 2022].

- Cms.gov. 2022. Self-Funded, Non-Federal Governmental Plans — CMS. [online] Available at: https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/nonfedgovplans [Accessed 9 April 2022].