Final Report: Predicting US Diabetes Prevalence

Springboard School of Data · Capstone Two

Author: Aisling Casey · Mentor: Tommy Blanchard · Date: July 2021

Contents

Overview

Background

Diabetes is a growing problem in the United States. The rate of diagnosed diabetes in American adults has increased more than 50% from 2000 and 2018, from a prevalence of 6.0% to 9.1%[i]. In 2017, about 1 in 7 dollars spent on healthcare in America was spent on treating diabetes and its complications[ii]. Knowing the state of diabetes at the local level informs public policy and can improve the efficacy of public health outreach.

The Centers for Disease Control (CDC) periodically releases the National Diabetes Statistics Report, which provides county level information on various diabetes related data, such as prevalence. While a report on the prevalence of diabetes at the national level was released for 2020, the last report with county level data was released for 2017[i].

Though it can afflict people of all backgrounds, developing type II diabetes is known to be affected by demographic factors such as race and income level[iii]. These factors can encourage behavior that make developing type II diabetes more probable. The National Diabetes Statistics report doesn't distinguish between type I and type II, though type I diabetics comprise a lesser share of the population - in 2016, the US diabetic population was roughly 85% type II diabetics, while the other 15% were type I[iv].

Objectives

Given that type II diabetes is influenced by demographic factors and comprises most diabetes cases, the primary objective of this project was to evaluate whether previous trends and new demographic data could be used to predict county-level diabetes prevalence in the United States. This model would be evaluated using the last diabetes prevalence data from 2017 as a validation set. If it proved more skillful than the baseline approach of using 2016's values as predictions for 2017, it could then be used to make predictions for the year 2018 & 2019. Secondary objectives of the project included determining which demographic factors are most important in predicting diabetes prevalence and understanding the burden of diabetes in all US counties over the last 15 years.

Results

An LSTM Deep Learning Network, accepting sequences 4 years in length, was selected. Using the 2017 data as a validation set, it yielded a mean absolute error of 0.78%, compared to a mean absolute error of 0.88% when using 2016's prevalence values as the prediction for 2017. This proves that changes in demographic data are useful in predicting changes in diabetes prevalence at the county level. Furthermore, the modeling showed that the socioeconomic factors, bachelor's degree prevalence and median income, are most important demographic factors in predicting Diabetes prevalence at the county level. Notebooks and the final model are available on the project repository on Github*.

Methods

Data Wrangling

The American Community Survey (ACS) compiles detailed estimates of demographic data in US communities. It releases one and five year estimates; one year estimates are only available for counties with populations over 65,000. The data is made available through the US Census Bureau.

Demographic data for 2006-2019 was pulled from the US Census Bureau API[v] using a custom function The variables pulled are detailed in figure 1 below. This data was used as independent variables in the predictive models and necessitated that only those counties with populations >65,000 be included in the model.

Feature Category	Detail	Type	Source	Availability
Diabetes	% Prevalence 20+ years	Target	CDC	-2004-2017 -All Counties
Population	# 20+ years	Explanatory	Census	-2006-2019 -Counties w/ population >65,000
Sex	# Total	Explanatory	Census	
Race	# Total (x6)	Explanatory	Census	
Age	# in 10 year bins	Explanatory	Census	
Education I	% Highschool or >	Explanatory	Census	
Education II	% Bachelor's or >	Explanatory	Census	
Income	$ Median household income	Explanatory	Census	

Figure 1 – Details of data used in predictive modeling

For each county, a FIPS state and county code is given. Taken together, these codes are unique identifiers for each county in the US. In all data sets, these codes are used to join data rows.

The diabetes prevalence data was manually downloaded as csv files from the United States Diabetes Surveillance System[vi], which is made available by the CDC. This data is currently available from 2004-2017, and reflects the prevalence of diabetes (type I & type II) in adults 20+ years of age. Unlike the American Community Survey there is no county population threshold for this data; of course, the smaller the population, the less accurate the estimates. Nonetheless, as an objective of this project was to understand the progression of diabetes in all US counites, prevalence data for less populous communities was used for data analytics and exploration.

Two more data sources were needed to understand the burden of diabetes across the United States: population data for each county for each year, regardless of county size, and an estimate of the fraction of the county 20+ years of age. As part of the Population Estimates Program (PEP), the US Census Bureau releases county population estimates for each year; this information was also queried from the Census API using a custom function. To get an estimate of the ratio of 20+ year old adults in each county, the ratio of each county in 2010 was gleamed from the County Intercensal Dataset for 2000-2010, which was downloaded as a csv from the Census website[vii].

Exploratory Data Analysis

All Counties – Analytics

After the 2010 age ratio was used to calculate the 20+ adult population in each US county for every year between 2004-2017, it was multiplied by the diabetes prevalence data to approximate the adult diabetic population in each county. The prevalence across regions could then be determined, as is show in figure 2. This graph reveals a steady increase across all regions, with the South having the greatest prevalence overall as well as the greatest increase, from 8.15% to 10.55%. The West has the lowest overall prevalence.

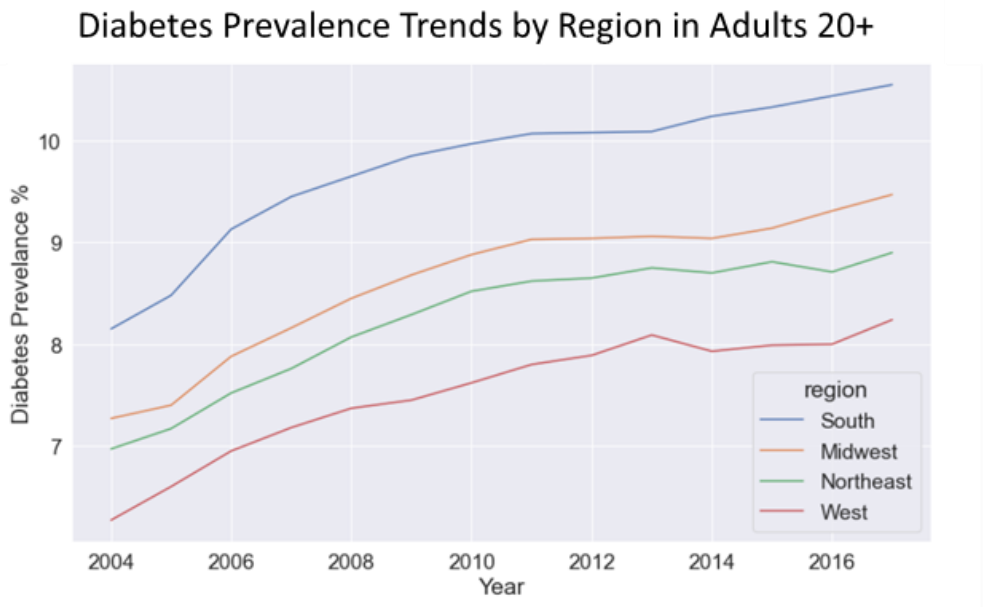

Figure 2 – Estimate of diabetes prevalence in US regions from 2004-2017, with data from all counties included

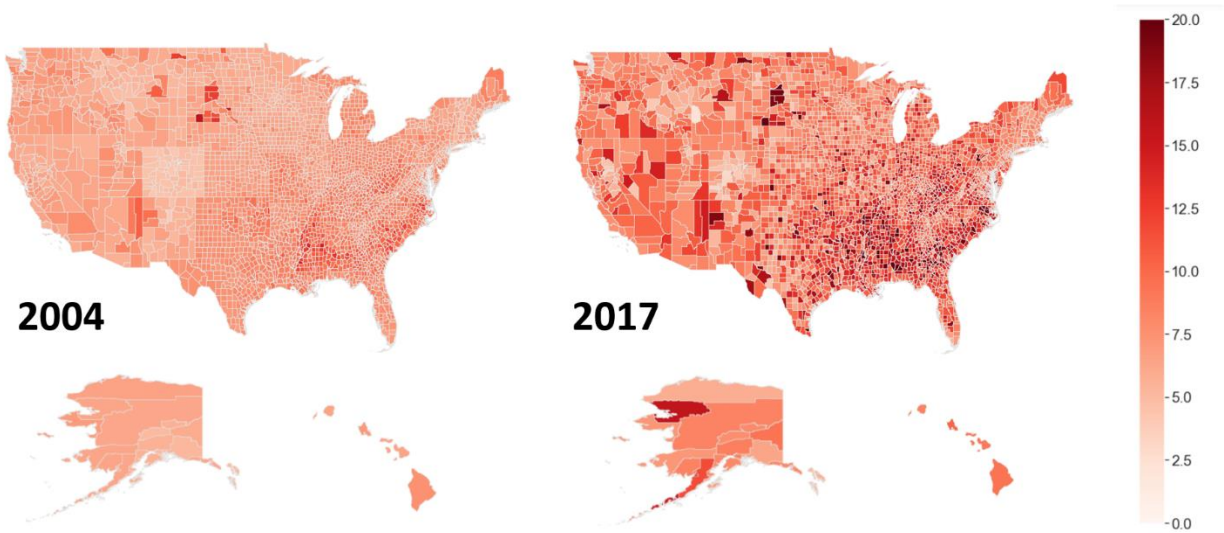

Figure 3 – Map of diabetes prevalence (%) 2004 and 2017, all counties. The same color scale is used for both maps

Figure 3 illustrates the change in diabetes prevalence in all counties from 2004 to 2017. It is a stark illustration of how diabetes prevalence has grown, particularly in Southern and Lower Midwestern counties, which tend towards being rural and less populous. The intersection of total population and diabetes prevalence is highlighted in figure 4 below, which displays the change in prevalence overtime, stratified by county population size. Counties with populations less than 100K see the greatest increase, from 7.8 to 11% in 13 years.

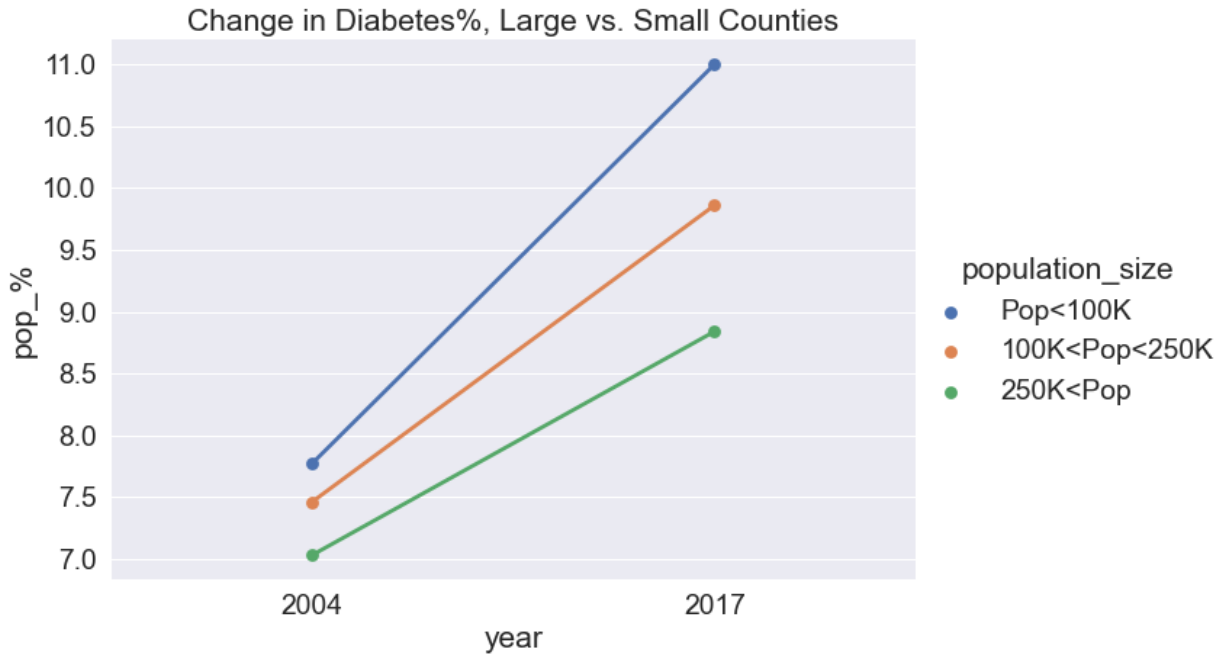

Figure 4 – Change in diabetes prevalence from 2004-2017, stratified by county population size

Highly Populated Counties – Modeling

Having explored diabetes prevalence in all counties, the rest of this report focuses on the diabetes prevalence and demographic variables of the more highly populated counties (i.e. >65K). This is the data that will be used to build a predictive model.

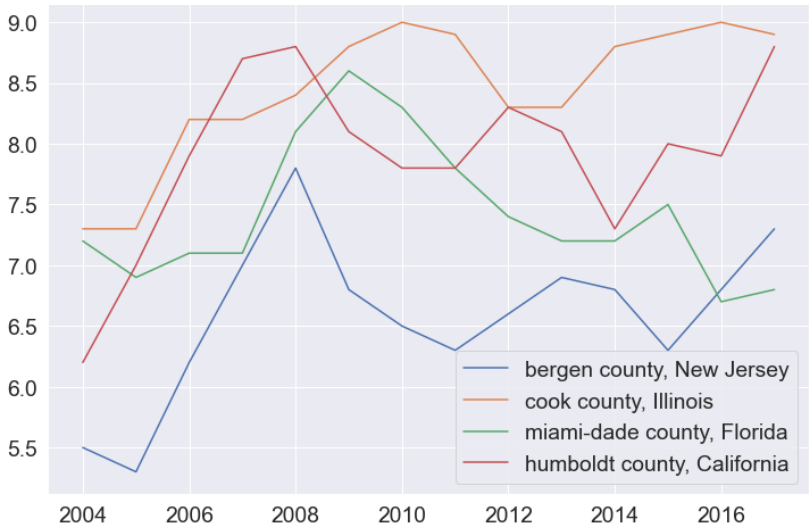

Figure 5 – Diabetes prevalence (%) overtime in 4 counties from different regions

Above, figure 5 shows the diabetes prevalence trend overtime in counties from the four different regions. The sharp peaks and dips in this graph illustrate that there is significant noise in the data. This is not surprising as it estimated from surveys of a fraction of the population; the same is true of the demographic data of the ACS data. The noise of both the dependent and independent variables will limit the predictive power of the model.

To understand the correlation between variables, the Pearson's correlation of the raw data was calculated. However, since many of the demographic variables were in terms of absolute counts per county (e.g. number of 25-34 year-olds, number of females), this yielded very high correlation between many variables. Thus, age, sex and race variables were changed from absolute counts to relative percentage of the population.

Furthermore, Pearson's correlation is more susceptible to outliers. Since there are some counties with a huge difference in some variables, Spearman's correlation seemed a better fit. Spearman's correlation uses the relative position, or rank of each variable to calculate correlation rather than value.

The Spearman's correlation of the relative features are shown in figure 6. It shows the male and female variables having an exactly -1 correlation, which makes sense; one of these variables (male) was later dropped from the dataset before modeling. It shows the older age brackets having high positive correlation with one another, as well as the socioeconomic (bachelor's %, high school % and median income) variables being highly positively correlated. No single demographic variable is highly correlated with the diabetes prevalence variable; this will be further explored in later graphs.

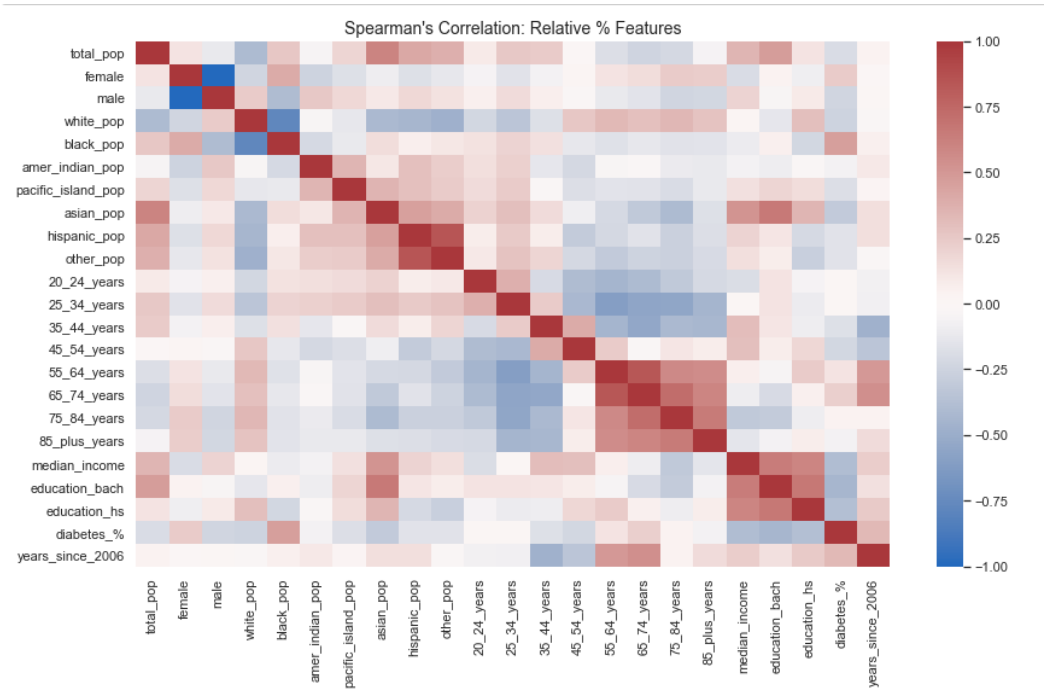

Figure 6 – Spearman's Correlation of all features. Note than age, sex and race variables are in terms of relative %

Figure 7 shows a scatter plot of education level and diabetes prevalence, with a line of best fit plotted as well. The percentage of the county with a bachelor's degree or more is shown in blue; a high school degree or more, orange. This graph illustrates that on the county level, there is a negative correlation between education level and diabetes prevalence.

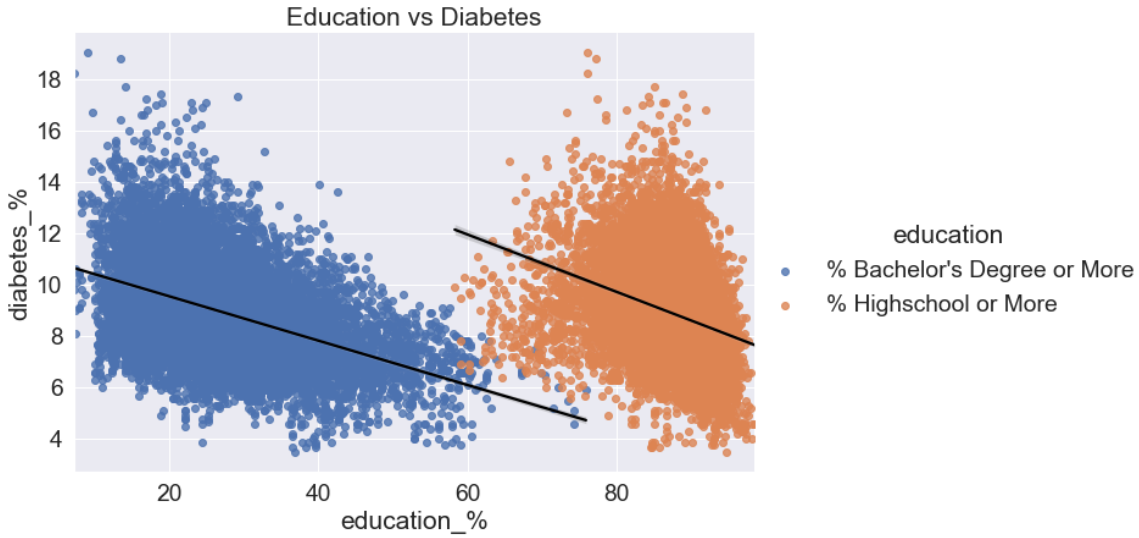

Figure 7 – Education vs. diabetes on the county level, 2004-2017

Relative percentage of a race in a county is plotted against diabetes prevalence in that county in figure 8. On the county level, diabetes prevalence goes down when the proportion of the while population goes up; the opposite is true for Black and American Indian populations. This agrees with the national prevalence by race; in 2018, 11.7% of the black population and 14.7% of the American Indian population suffered from diabetes, compared to 7.5% of the white population[viii].

The graphs of the Hispanic and Asian populations are an illustration of Simpson's paradox. While at the national level, these groups had a diabetes prevalence of 9.2% and 12.5% in 2018[viii], which was higher than the average prevalence, on the county level the opposite trend is shown. As the share of these populations in a county goes up, the diabetes prevalence trends downwards. This is an illustration of how correlation does not imply causation, and that further analysis is needed to determine which factors are most important in predicting diabetes.

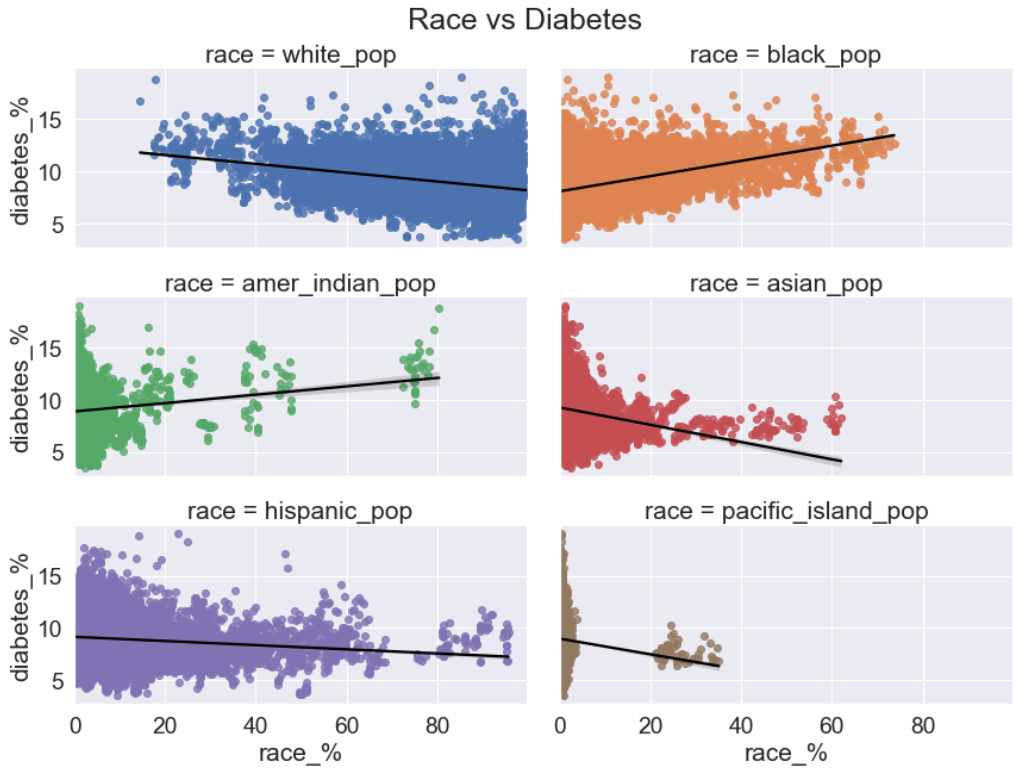

Figure 8 – Race vs. diabetes on the county level, 2004-2017

Preprocessing & Modeling Considerations

The following general preprocessing steps were taken to prepare the data for all models:

- Convert absolute counts of age, sex and race in relative %
- Drop counties with population <65K for >3 years
    - Back & forward fill remaining values
    - Years-> 'years_since_2006'
- Dummy-encode region
- Log transform to normalize skewed variables (see figure 9)
- Min & max scaling

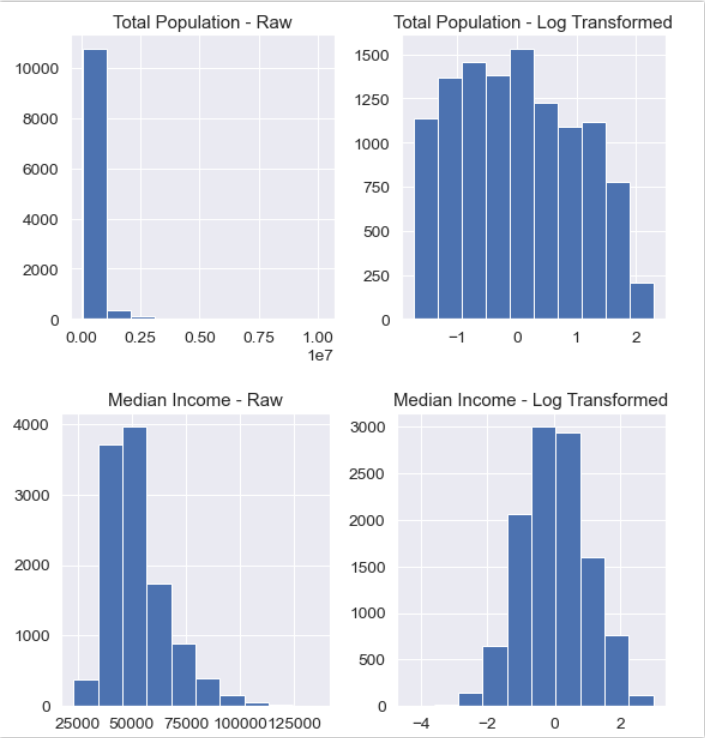

Figure 9 – Total population and median income distributions before and after log transformation

There were a few considerations in deciding how to model the data. Prevalence estimates for each county were desired, but there was a limited data pool– only 12 time points where the demographic data and diabetes prevalence data overlapped for each county. It was important to capture the county level trends, but this could not be accomplished with an autoregressive model as there were not enough data points. Furthermore, diabetes prevalence is trending upwards, so the model would need to be able to make out of sample predictions, ruling out tree-based models. Finally, it was important to understand which features most affected prediction.

Years Since 2006	Diabetes %	Diabetes % (t-1)	Diabetes % (t-2)	Diabetes % (t-3)
1	8.1	7.8	7.2	7.2
2	8.6	8.1	7.8	7.2
3	9.9	8.6	8.1	7.8
4	10.2	9.9	8.6	8.1
5	9.8	10.2	9.9	8.6

To capture trends in each county, the diabetes prevalence value in the one, two and three years preceding each data point was added to the dataset for a linear regression and a feedforward neural net model. This feature engineering is depicted in figure 10, where each row represents the data for one year in one county. Next, the data was split into training data (2006-2016) and testing data (2017) for the linear regression and feedforward neural net models.

Figure 10 – Adding the three previous year's (t-n) diabetes prevalence value to each column

Multivariable LSTM networks with various sequence length inputs were also created. LSTM networks models are intended to work with sequences of data, as they can learn what input information is important for future predictions and what should be forgotten.

To use these models, the data shape had to be changed. This is depicted in figure 11, which shows the training features input for an LSTM network of sequence length n=4 being 6456 sequences, 25 columns in width and 4 rows in length. These training sequences would be fed into the model fitting algorithm with a 1-dimensional array of length 6456, corresponding to the diabetes prevalence value associated with each of the 6456 feature sequences. This data would then be validated with a testing feature array of size (807, 4, 25) and a dependent variable array of (807), representing the data from 807 counties for the year 2017.

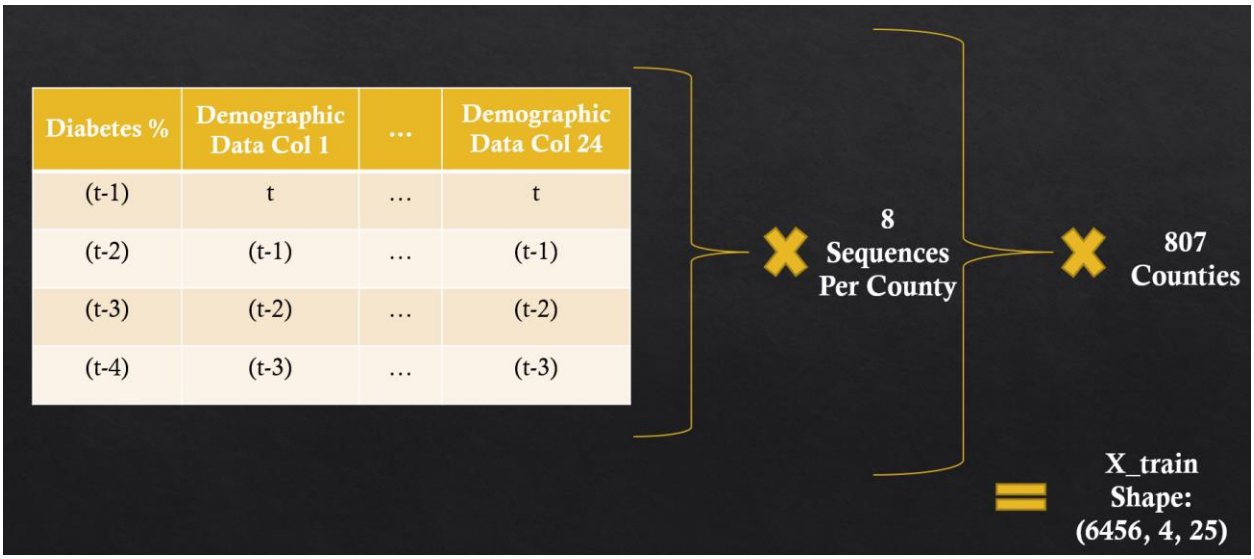

Figure 11 – Structure of LSTM n=4 training features

Modeling & Performance Metrics

Each model's performance was validated using the 2017 diabetes values for each county. The mean absolute error (MAE) between the predicted and actual values was determined and compared to the baseline approach of using 2016's values as the values for 2017. The baseline approach yielded an MAE of 0.876.

Linear Regression

The linear regression model was created using sklearn's Ridge regression class. The optimal alpha value was determined to be 1. This hyperparameter yielded the model with the lowest MAE score of 0.801, an 8.5% improvement over the baseline approach.

Additionally, a linear regression model with no regularization was created using the statsmodels package OLS to calculate the model coefficients, their p-values and confidence intervals. This was done to understand feature importance rather than to build a predictive model. Regularization was not implemented because there currently is no summary statistics for regularized linear models in the statsmodels package. The coefficient values for both models are shown in figure 14 in the discussion section.

Feedforward Neural Net

A feedforward neural net was created using three Dense layers, the final layer returning one value and having no activation function to create a regression model. The first two layers had 'relu' activation functions, used MAE as a loss function and the 'adam' optimizer with standard parameters. The model fit call was configured with 100 epochs, but an early stop callback was added to prevent any further models from being fit after 10 epochs in a row failed to yield any MAE improvement. The epoch that yielded the lowest MAE was saved as the final model.

Due to the non-deterministic nature of neural net creation, 50 of these models were created to get a sense of true performance of this model configuration. This is illustrated in figure 12, where it is shown that 50 feedforward neural nets had an average MAE of 0.806.

Multivariable LSTM Network

LSTM networks with input sequence lengths of 3, 4, 5 & 6 years were created. The first layer of these networks were LSTM layers with the default calls, such as activation='tanh', and had an 'adam' optimizer as well. The second and final layer had no activation function and a single output. These models also had an early stop call back with a patience of 10 epochs and saved the model from the epoch with the lowest MAE.

LSTM models of sequence length n=3 tended to have MAE of over 0.8, so it was not experimented with further. On the other hand, models of sequence length n=6 had large swings between MAEs of 0.9 to 0.75 – indicative of this model's tendency to overfit. 50 iterations of the other two models, sequence length n=4 and n=5, were generated to assess the stability of their performance.

It was found that LSTM networks of length n=4 had a lower mean MAE than the feedforward neural net at 0.786 and a lower standard deviation in MAE than the n=5 network. This shows that this model has a better balance between bias and overfitting than the other models, and thus an LSTM network of length n=4 was chosen as for the final model. The distributions of the MAE are shown in the graph below.

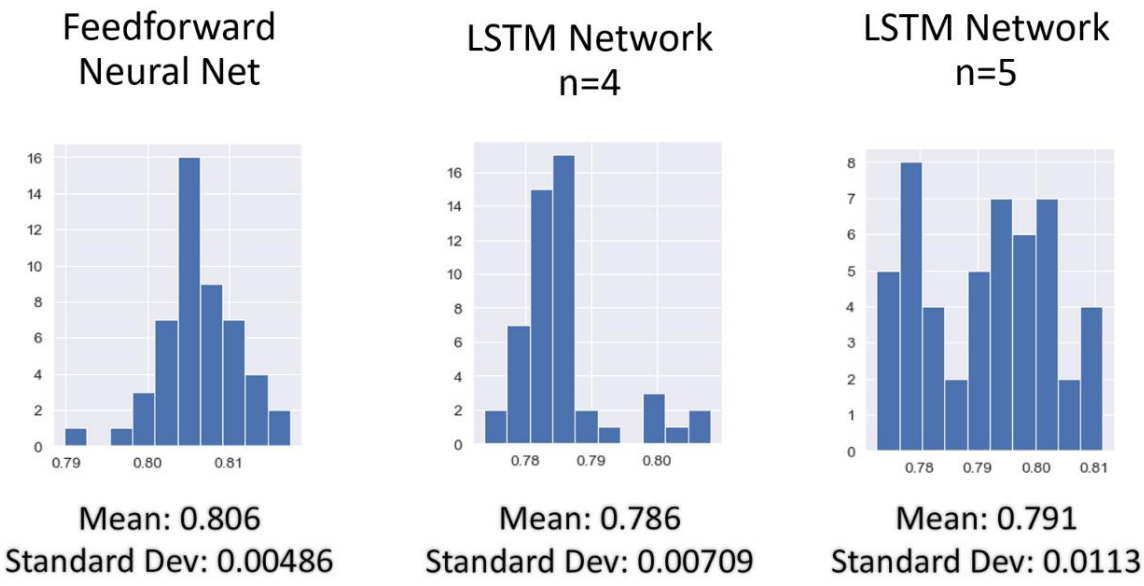

Figure 12 – Histograms of MAEs from 50 iterations of each model

The final model selected was an LSTM n=4 model with an MAE of 0.784. This is a 10.5% improvement on the MAE of the baseline approach.

To further explore the performance difference between these two approaches, the counties were broken up into quartiles according to total population. Then the absolute difference between each prediction approach and the actual value for 2017 diabetes prevalence was bootstrapped. The mean of these differences was calculated to create a bootstrapped, quartile mean absolute error. This process was repeated 50 times, and the results are shown in figure 13. For each quartile, the final model yielded a lower MAE on average, and had a slightly lower standard deviation.

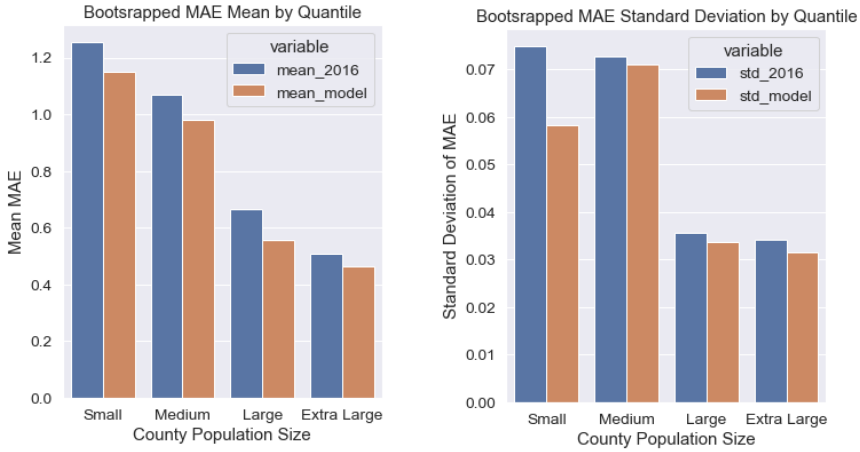

Figure 13 - Mean and standard deviation of bootstrapped MAE for each population quartile

Discussion

The primary objective of this project was to determine if changes in demographic data could improve upon the approach of predicting diabetes prevalence values at the county level using only the previous year's data. The final multivariable LSTM model does that, improving upon the MAE by 11%. This is a marginal increase; however, given the noise in the data (both the features and dependent variable are approximated by data), the model performs reasonably well. On a more fundamental level, the model reflects that demographic factors do play a role in diabetes prevalence.

Another objective was to understand feature importance in diabetes prediction. In figure 14, the results from linear regression are shown from two different model libraries, one regularized and one not. While they show slightly different results, they both indicate that, behind previous years' diabetes values, socioeconomic indicators such as bachelors' degree % and median income are some of the most important variables in predicting diabetes prevalence. Race also tends to be more important than age or sex in predicting diabetes prevalence. Furthermore, most features are found to be statistically significant by the non-regularized model.

This project has shed light on the growing diabetes burden in the United States. It is a particularly unequal disease, afflicting certain minority races, lower income individuals and rural areas disproportionality, to the point that diabetes prevalence can be predicted using this demographic data. Knowing the current state of diabetes prevalence across the country is a crucial aspect of public health outreach, and the final model from this project can be used for predicting this prevalence.

	Feature	Coefficient
24	diabetes_1_year_past	[12.88686701226523]
18	education_bach	[-1.362412827915762]
26	diabetes_3_years_past	[-1.348861333363724]
7	hispanic_pop	[-0.8407621200811116]
25	diabetes_2_years_past	[0.779002301608746]
4	amer_indian_pop	[0.6361382290824514]
17	median_income	[-0.6350155226495989]
6	asian_pop	[0.5623975488746945]
3	black_pop	[0.5252997130072462]
20	years_since_2006	[0.4730568068488451]
5	pacific_island_pop	[-0.4190678348761399]
11	35_44_years	[0.40367914023024004]
12	45_54_years	[-0.33613467792998747]
16	85_plus_years	[-0.3278575438893192]
19	education_hs	[-0.3210219349622421]
9	20_24_years	[-0.32060929068264743]
10	25_34_years	[-0.2978144330078545]
14	65_74_years	[-0.2546054848985859]
13	55_64_years	[-0.1996583064988869]
23	region_West	[-0.1895324394800737]
8	other_pop	[0.15494335566230444]
15	75_84_years	[-0.148346244696043]
1	female	[0.11437752044035952]
22	region_South	[0.0902446546714452 4]
2	white_pop	[-0.057222732303445295]
21	region_Northeast	[0.028110099476753404]
0	total_pop	[-0.021488790103826413]

	p	coef	2.5%	97.5%
diabetes_1_year_past	0	14.0418	13.645	14.439
const	0	5.1329	4.029	6.237
diabetes_3_years_past	0	-1.4072	-1.728	-1.087
education_bach	0	-1.2907	-1.502	-1.079
hispanic_pop	0	-0.8132	-1.026	-0.601
median_income	0	-0.6391	-0.873	-0.405
years_since_2006	0	0.478	0.353	0.603
region_West	0	-0.184	-0.244	-0.124
region_South	0	0.0861	0.039	0.133
asian_pop	0.002	1.1804	0.438	1.923
amer_indian_pop	0.004	1.3698	0.446	2.294
85_plus_years	0.005	-0.3147	-0.533	-0.097
black_pop	0.006	1.2269	0.354	2.1
education_hs	0.008	-0.3059	-0.532	-0.079
25_34_years	0.009	-0.3573	-0.625	-0.09
35_44_years	0.016	0.3837	0.07	0.697
45_54_years	0.017	-0.3565	-0.65	-0.063
pacific_island_pop	0.021	-0.4929	-0.911	-0.075
20_24_years	0.028	-0.3611	-0.684	-0.039
other_pop	0.038	0.5574	0.031	1.084
white_pop	0.122	0.7927	-0.213	1.799
55_64_years	0.153	-0.1924	-0.456	0.071
region_Northeast	0.218	0.0309	-0.018	0.08
65_74_years	0.267	-0.346	-0.957	0.265
female	0.547	0.0817	-0.184	0.348
75_84_years	0.584	-0.1126	-0.516	0.291
total_pop	0.639	-0.0198	-0.102	0.063

Figure 14 – Coefficients for the linear models; an L2 regularized sklearn model (left) ordered by absolute coefficient value and a non-regularized OLS model from statsmodels (right) ordered by p-value, then absolute coefficient value

References

*Project GitHub Repository

https://github.com/Aisling-C/Springboard/tree/main/CapstoneProject2/DiabetesPrevalence

Tableau Dashboard

https://public.tableau.com/app/profile/aisling.casey/viz/Diabetes_Prevalence/USDiabetesPrevelance2004-2017

Citations

[i] "U.S. Diabetes Surveillance System." Centers for Disease Control and Prevention. Accessed June 19, 2021. https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.

[ii] "The Cost of Diabetes." American Diabetes Association. Accessed June 19, 2021. https://www.diabetes.org/resources/statistics/cost-diabetes.

[iii] Hill, James P. et al. 'Scientific Statement: Socioecological Determinants of Prediabetes and Type 2 Diabetes.' Diabetes Care 2013 Aug; 36(8): 2430-2439. DOI: 10.2337/dc13-1161. Accessed 29 July 2021.

[iv] Bullard, Kai M. et al. 'Prevalence of Diagnosed Diabetes in Adults by Diabetes Type — United States, 2016.' MMWR Morb Mortal Wkly Rep 2018; 67:359–361. DOI: 10.15585/mmwr.mm6712a2. Accessed July 30 2021.

[v] "American Community Survey Data." United States Census Bureau. Accessed July 10, 2021. https://www.census.gov/programs-surveys/acs/data.html.

[vi] "Diagnosed Diabetes." United Stats Diabetes Surveillance System, Centers for Disease Control and Prevention. Accessed June 23, 2021. https://gis.cdc.gov/grasp/diabetes/diabetesatlas.html#

[vii] "County Intercensal Datasets: 2000-2010." United States Census Bureau. Accessed July 10, 2021. https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-counties.html

[viii] "Statistics about Diabetes." American Diabetes Association. Accessed July 27, 2021. https://www.diabetes.org/resources/statistics/statistics-about-diabetes.