MACHINE LEARNING CLASSIFICATION MODEL

DIABETES IN U.S. COUNTIES

June 2020

Francis Morales

PROJECT OBJECTIVE

Determine the level of population with **Diabetes** in counties of the US to guide **retailers** on determining their **sugar conscious** merchandise **stock**.

DIABETES
IN THE UNITED STATES

34.2 Million

1 in every **10** people has Diabetes

7th

leading cause of **death**

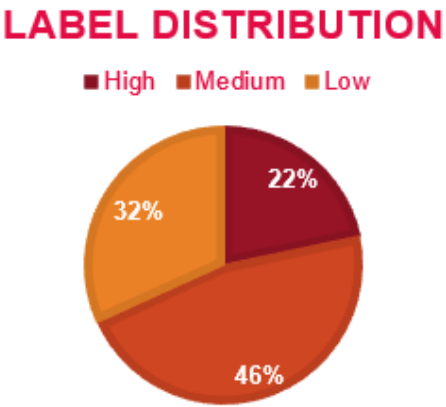
1.5 Million

people 18 years old or older **diagnosed** with diabetes in **2018**

DIABETES MANAGEMENT

DATASET

- **820** counties from 48 states (**26%** of counties)
- Excluded Hawaii and Alaska from study
- **50** different features: economic features, demographics, geographic coordinates, weather components & food access
- **7** Data Sources: Web scraping, API & data download
- Diabetes labels – by population with diabetes percentage
  - **Low:** 0% to 8.5%
  - **Medium:** 8.6% to 11.6%
  - **High:** more than 11.6%



United States Census Bureau

Zillow

menuism
RESTAURANT AND DISH REVIEWS

WIKIPEDIA
The Free Encyclopedia

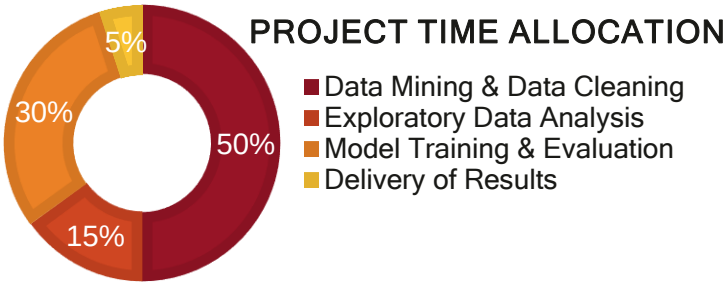
USDA United States Department of Agriculture

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

NOAA
NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

METHODOLOGY

Data Mining & Data Cleaning » Exploratory Data Analysis » Model Training & Evaluation » Delivery of Results



Trained models using Pipelines & Grid search:

- Random Forest
- **SVM**
- K-Nearest Neighbors
- Adaboost/Gradient Boosting/ XGBClassifier

MODEL OVERVIEW

A Support Vector Machine Classification Model that predicts the level of population with Diabetes of a county with **70.7%** accuracy.

Classifier	Accuracy	Precision (wgt avg)	Recall (wgt avg)
Random Forest	0.6747	0.67	0.67
KNN	0.6422	0.65	0.64
SVM	0.7073	0.71	0.71
XGBoost	0.7073	0.70	0.71

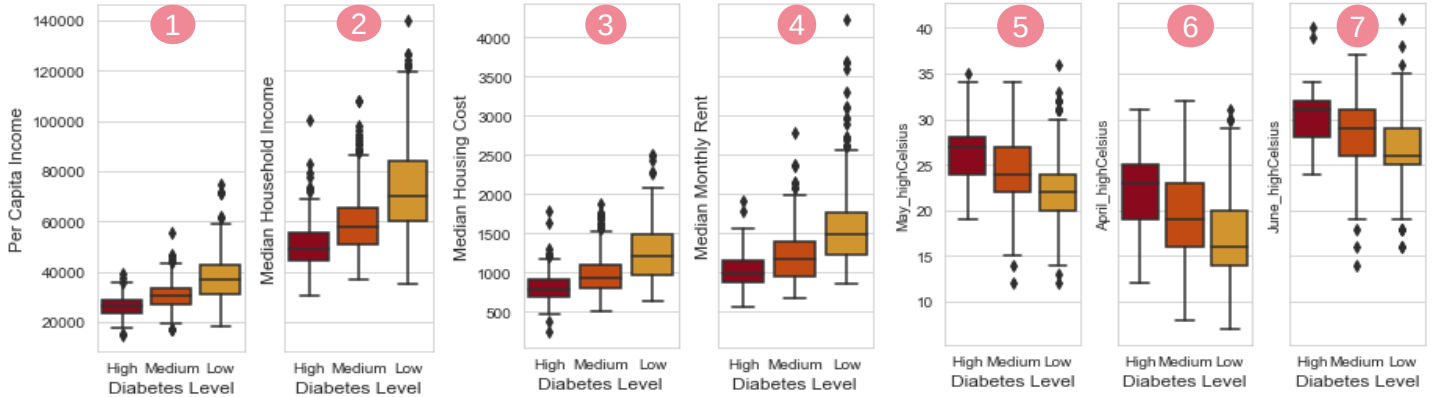
MOST INFLUENTIAL FEATURES

Per Capita Income
Median Household Income
Median Housing Costs
Monthly Rent
May High Temperature
April High Temperature
June High Temperature

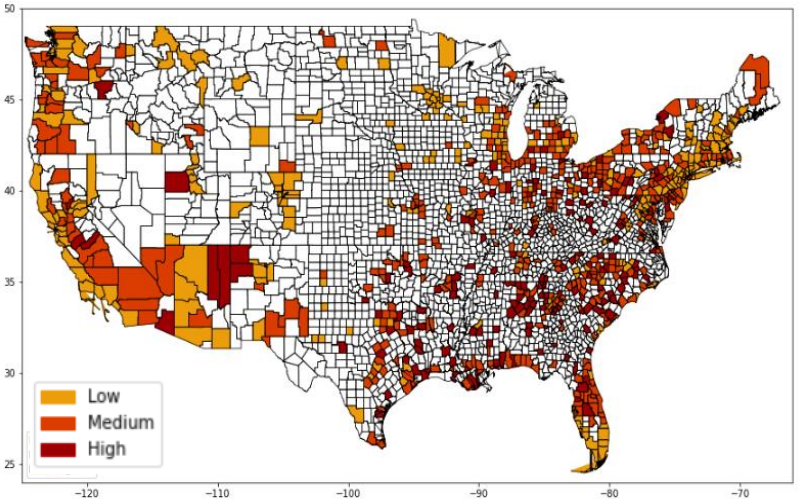
More income, less diabetes

Higher living costs, less diabetes

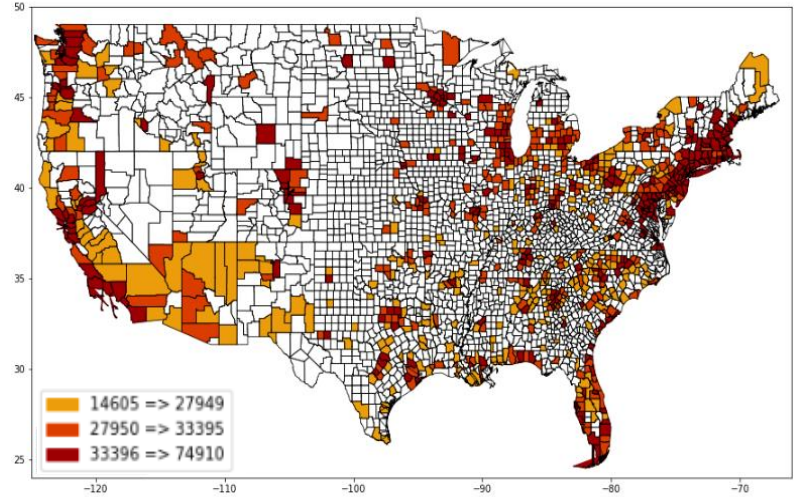
Milder Spring temperatures, less diabetes



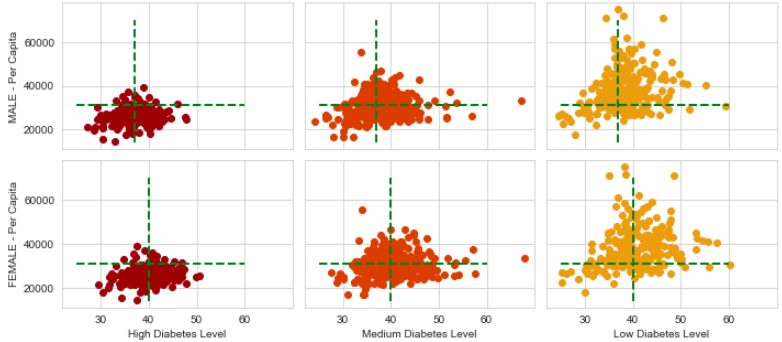
DIABETES LEVEL



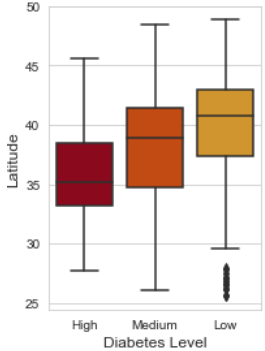
PER CAPITA

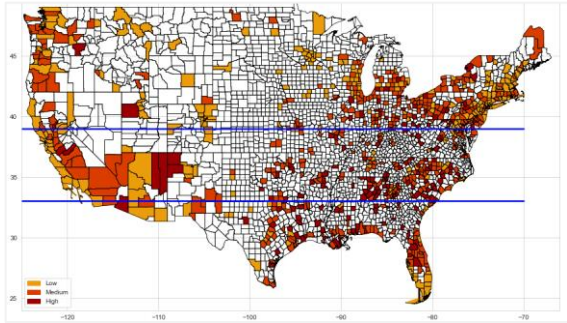


PER CAPITA INCOME VS MEDIAN AGE



LATITIDE





RECOMMENDATIONS

Use the Machine Learning classification model to identify the level of diabetes of your store's location

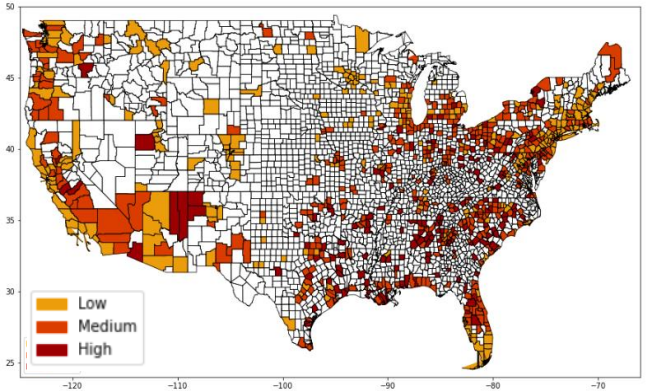
Carry affordable diabetic friendly items in stock in counties with High levels of diabetes

Invest in labeling diabetic-friendly items in counties with High levels of diabetes as more than 12% of the customers that walk in will be interested in that selection

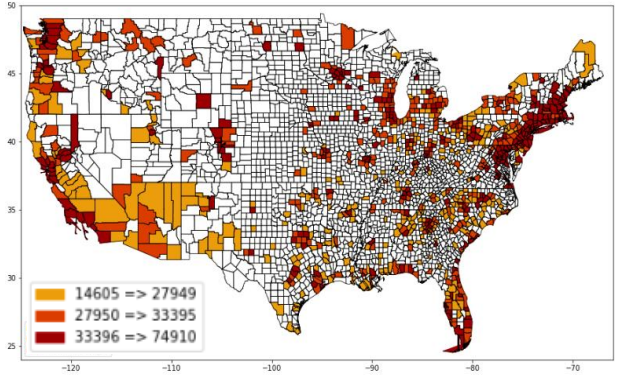
Thanks!

APPENDIX A

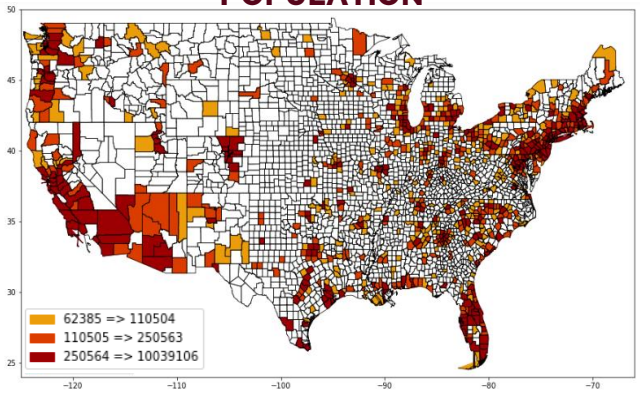
DIABETES LEVEL



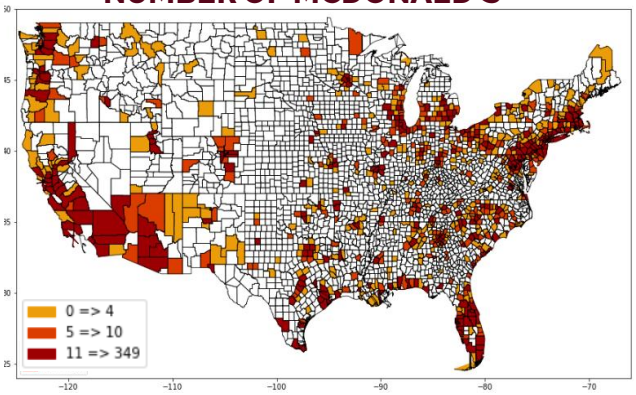
PER CAPITA



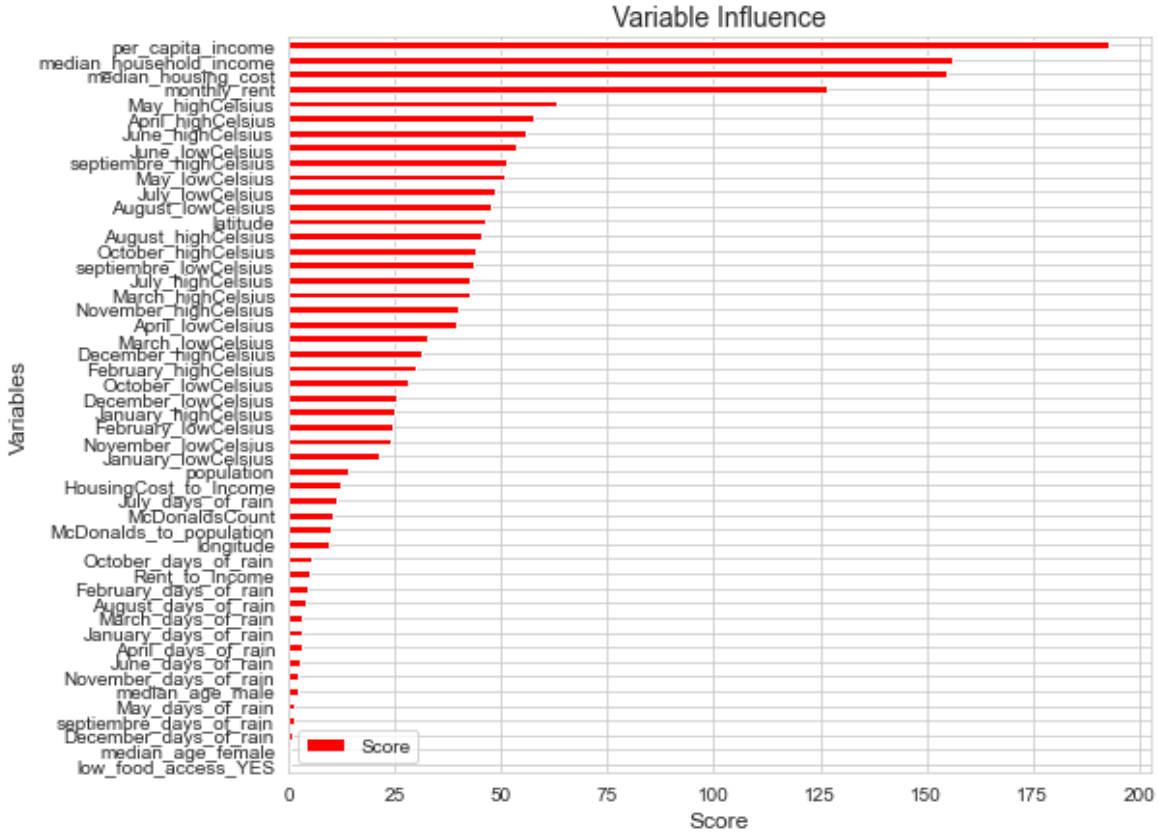
POPULATION



NUMBER OF MCDONALD'S



APPENDIX B


Variable Influence
per_capita_income
median_household_income
median_housing_cost
monthly_rent
May_highCelsius
April_highCelsius
June_highCelsius
June_lowCelsius
septiembre_highCelsius
May_lowCelsius
July_lowCelsius
August_lowCelsius
latitude
August_highCelsius
October_highCelsius
septiembre_lowCelsius
July_highCelsius
March_highCelsius
November_highCelsius
April_lowCelsius
March_lowCelsius
December_highCelsius
February_highCelsius
October_lowCelsius
December_lowCelsius
January_highCelsius
February_lowCelsius
November_lowCelsius
January_lowCelsius
population
HousingCost_to_Income
July_days_of_rain
McDonaldsCount
McDonalds_to_population
longitude
October_days_of_rain
Rent_to_Income
February_days_of_rain
August_days_of_rain
March_days_of_rain
January_days_of_rain
April_days_of_rain
June_days_of_rain
November_days_of_rain
median_age_male
May_days_of_rain
septiembre_days_of_rain
December_days_of_rain
median_age_female
low_food_access_YES
Variables
Score
0
25
50
75
100
125
150
175
200
Score