The Mean Squares

Diabetes Risk Analysis

DATE

23rd February , 2018

© 2017 DMI CONFIDENTIAL & PROPRIETARY

DMI

DRIVING MOBILE INNOVATION

Diabetes Risk Prediction and Fitness Recommendation Model

Objective :- We will be developing model(s) to predict the chances of a person getting Diabetes based on certain parameters like weight , bmi , habits etc. and then recommending the ideal BMI and few other parameters within their lifestyle segment to reduce the probability.

Data Source :- Health Survey Data Collected by Centre of Disease Control (CDC) United States

Data Volume :- Survey data from approx. 4 lakhs participants (Approx. 350 MB csv file)

Approach:-

- Determine the key parameters/metrics which are correlated to probability of having Diabetes in a person
- Develop User Segments based on health and behavior metrics
- Create look-alike process to provide recommendation specific to the segment a new user matches with based on the clustering parameters

SOLUTION APPROACH

Data Extraction and Preparation

- Converting SAS data set into R data frame
- Identifying variables of interest
- Data conversion , cleansing and missing value treatment

Logistic Regression For Finding Factors Driving Diabetes Probability

- Iterative model creation using Logistic Regression
- Develop confusion matrix for model confidence assessment
- Predict for test data set

Clustering Analysis For Recommendations

- Normalize Clustering variables and determine number of Clusters needed
- Develop K means clustering model
- Profile Clustering Variables

Shiny App Development For Results

- Used the Shiny package to develop interactive user interface for Regression Model and Clustering sepreately

Logistic Regression Output

Test Data Validation

Prediction / Actual	0	1
0	6936	2613
1	2660	7030

% False Positive :- 13.8%

% False Negatives :- 13.5 %

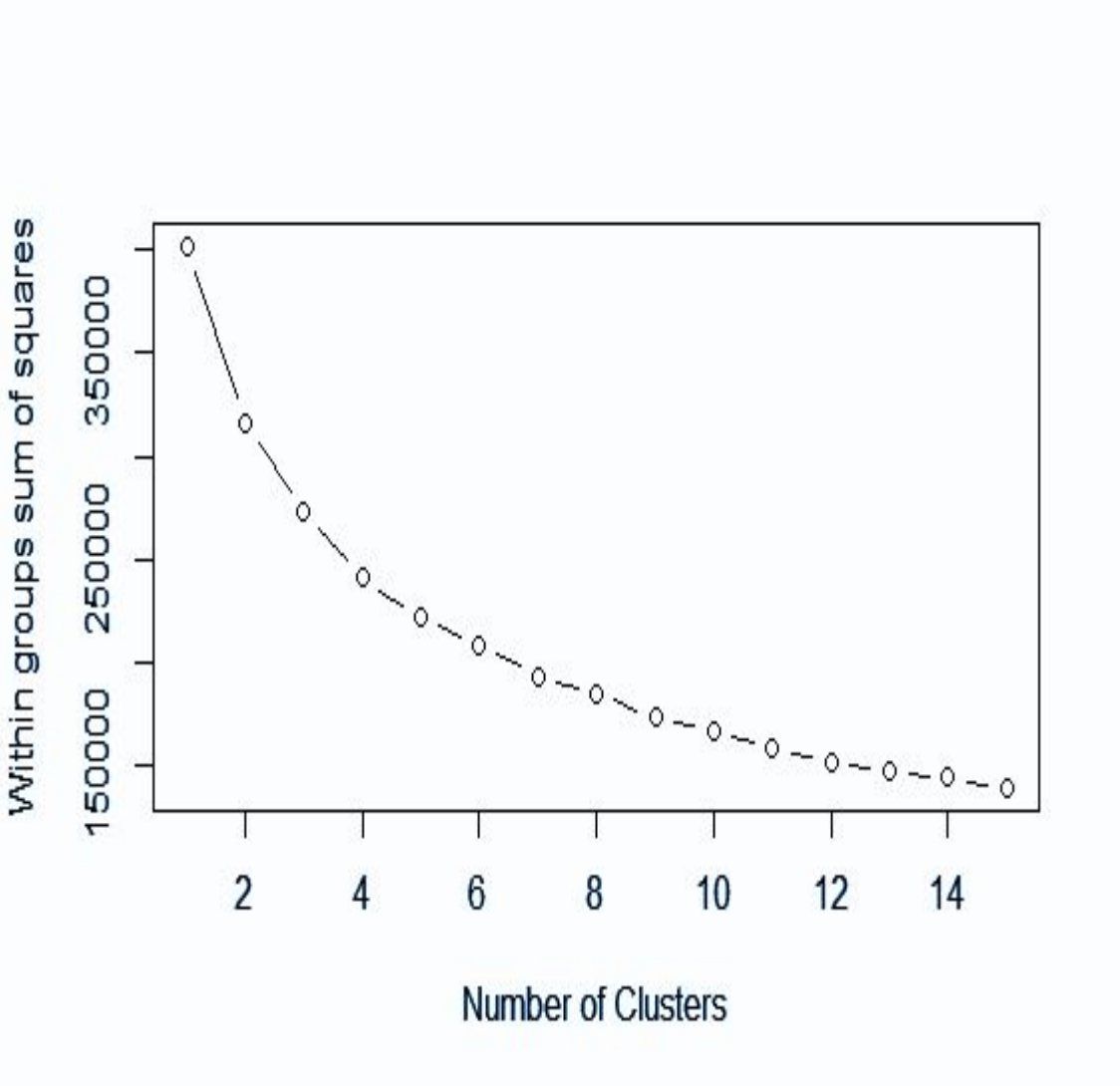
Model Accuracy :- ~75%

Genetics plays a very strong and important role in determining diabetic probability in a person. If we can include that data in the model, the accuracy will go up.

User Segmentation and Recommendations

- We have 6 different clusters based on Height , weight , age , sleeping hours and household income.
-

Cluster	% Users
1	15 %
2	22 %
3	13 %
4	19 %
5	17 %
6	14 %



- We have used lifestyle variables like Smoking Habits , Alcohol Consumption , Sleeping Hours and BMI for recommendation for leading a healthy life within a user segment.

Working Prototype Demo

What Next?

- Integration of Probability Prediction and Recommendation Engine together in one interface to create a seamless experience.
- Include more data elements in the models such biometric data (haemoglobin , feet check , blood sugar etc) , family history , lifestyle variables etc to improve the model's accuracy and robustness.
- Develop mobile/web apps for people to self monitor their diabetic risks based on these parameters continuously and take preventive actions or medical consultations as needed.
- Start ingesting data from connected fitness devices to improve the coverage of the model.
- Include some geographical and ethnic variables in the model to tailor it for different geography and targeted user base.

  Ex:- People living in Northern Climates have higher risk for Type 1 diabetes.