Paper 2806-2018

Demographic & Lifestyle Factors in Predicting Chronic Diseases

Asish Satpathy, UCR School of Business, University of California, Riverside, CA

ABSTRACT

Chronic diseases such as diabetes and hypertension are widely prevalent in the world and are responsible for a significant number of deaths each year. In addition, treatments for such chronic diseases account for an extremely high healthcare cost. Research has shown that these diseases can be proactively managed and prevented while lowering the healthcare cost. Using SAS® Analytics, we have mined a data sample of ten million consumers with over 1000 demographic and lifestyle attributes who have self-reported their chronic disease condition. We developed a model from a subset of demographic and lifestyle factors to demonstrate the accuracy of prediction of chronic diseases. The objective of this study is to understand if the currently accessible large volume of consumer data is useful in identifying top social factors which are associated with such diseases. When used with patient clinical data that can be beneficial in devising preventive treatment plans and thereby reducing healthcare costs.

Note: This is purely a data driven study with no input from experts in the healthcare field. We believe our study will generate interest in the healthcare community.

INTRODUCTION

According to the National Diabetes Statistics Report, 2017 [1], an estimated 30.3 million people–or 9.4% of the U.S. population–had diabetes in 2015. A further breakdown of that number reveals that the percentage of adults with diabetes increased with age, reaching a high of 25.2% among those aged 65 years or older. This drastic increase in the incidence of diabetes (Figure-1) could be attributed to changes in lifestyle during the last century for certain demographics, a condition termed as a “metabolic syndrome”.

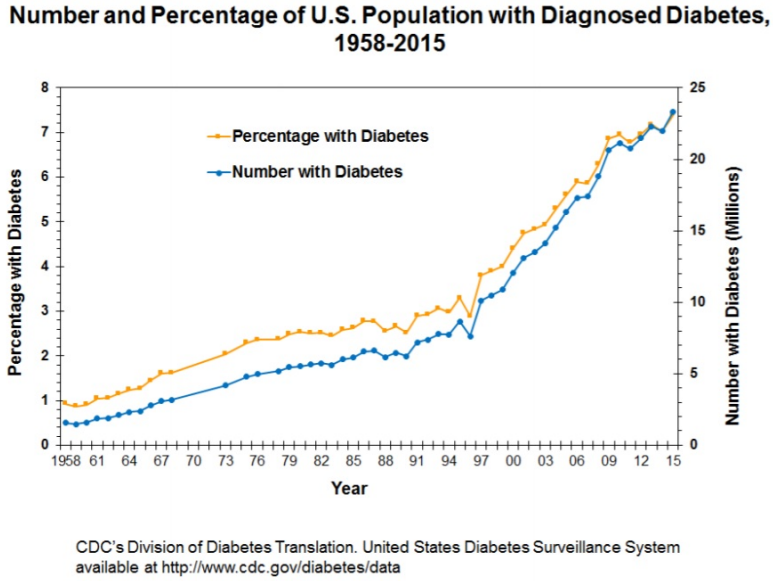


Figure 1. The rise of diabetes prevalence in the U.S.

Past studies to establish any associations between patients' demography/lifestyle and chronic diseases in general have not been conclusive. A typical clinical study with patients suffering from chronic diseases would likely include results of direct data collection that may contain some bias from the patients' willingness to share their demographic and lifestyle information. In this paper we take a different approach to determining such a possible association.

We have used data from over 10 million consumers (consumer and patient will be used interchangeably throughout this paper), collected by a large marketing firm for a marketing campaign study, to model possible associations of consumers' demography and lifestyle with their diabetic conditions. The disease information in the data has been self-reported by the consumers in the database. Our goal for this study is to develop answers to the following research questions:

(1) What model would predict consumers' diabetic conditions based upon their demography and lifestyle data?

(2) Can we rely on available consumer data with social behavior to predict chronic diseases?

(3) What will be the implication of our study regarding the proactive management of chronic diseases such as diabetes?

METHODS

According to the Center of Disease Control (CDC), diabetes has been on the rise in the US for the last several decades (Figure-1) and a serious preventive program is needed to stop this trend. A snapshot of the prevalence of diabetes across the U.S. states is shown in Figure-2. Since the severity of this disease is location-dependent, and demography and lifestyle of those patients are also location-dependent, we hypothesize that an association could exist between a consumer's demography/lifestyle and their chronic disease condition.

We hope to develop a model to confirm or refute such an association using survey response data from over one million consumers in a sample from the state of Texas.

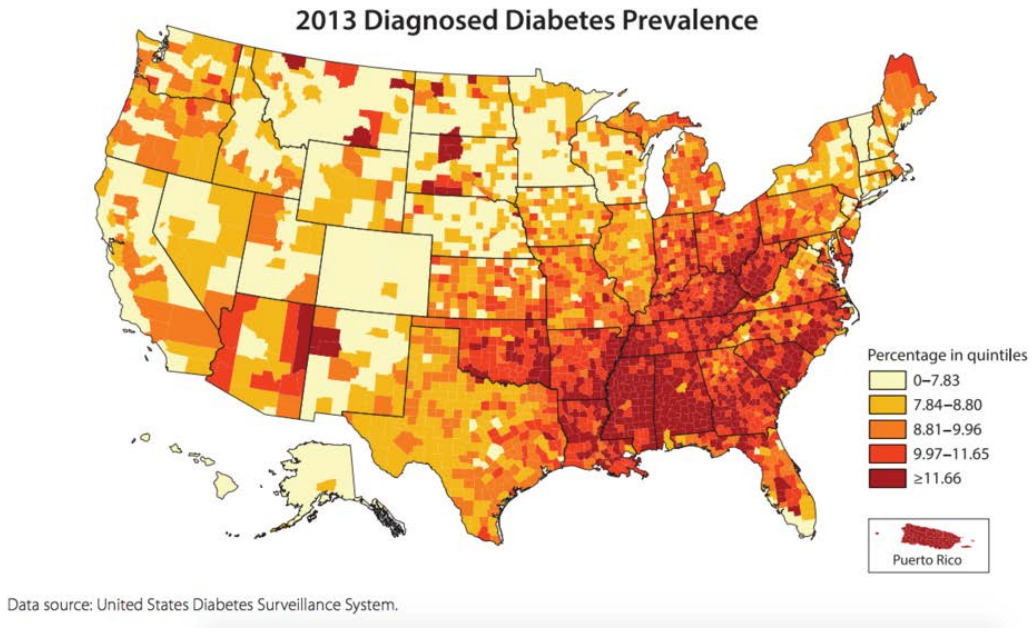


Figure 2: County level estimates of diagnosed diabetes among adults (aged > 20 years) in the U.S. (2013). Source: Center for Disease Control (cdc.gov).

DATA EXPLORATION

Over one million records of data from the state of Texas were available from a marketing firm with several key categories that uniquely identified patients for our study. BASE SAS® procedures were used to clean up, impute, and visualize the data. Table-1 shows a snapshot from over 1000 such characteristics of consumer data used in the analysis who either have or have not reported on their existing chronic illnesses. The lifestyle variables, which form a dominant fraction among the characteristics, each have a total of 99 levels, 1 being most likely and 99 being least likely a person would belong to that category. For this study, we selected a sample of 30,000 consumer records with self-reported diabetes (signal). We also collected a mutually exclusive data sample, derived from the same population, of about 50,000 consumers with other minor illnesses such as anxiety, allergy, osteoporosis, hearing loss, etc., who have not reported any signs of diabetes or other chronic illnesses. We treated this as our background sample for the purpose of model-building. Both the signal and background samples were obtained by simple random sampling from their respective populations.

Demography	**Census Data**	**Interests**	**Lifestyle**
Gender	Population	Music	Travel
Age	Households size	Sports	Online Preference
Education	Per Capita Income	Hobbies	Deposit Customer

Property	**Finance**	**Consumer**	**Credit**
Dwelling Type	Income	Channel Preference	Credit Score
Square Footage	Loans	Number of Purchases	Open Trades
Year Built	Credit Card Type	Social Media Users	Risk Predictor

Table 1. A sample of patient profiles in our data who may or may not have diabetes.

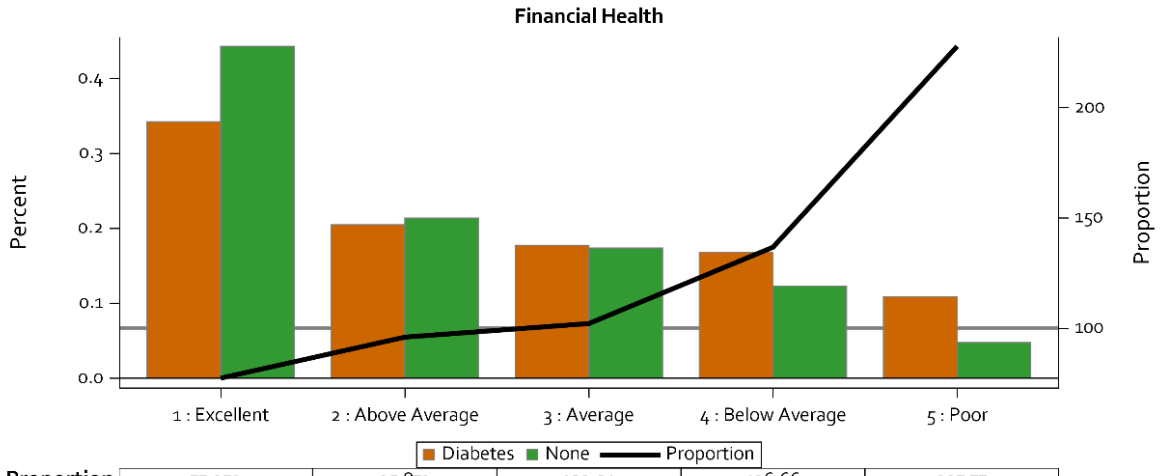
Upon exploring the data, we found several demography and lifestyle related predictor variables in the data sample showing significant differences between the diabetic and non-diabetic population. In order to visualize such differences, we computed *proportions*, which is a ratio of percentage of diabetic and non-diabetic consumers, for each bin of dependent variables. For a certain category of social behavior, when this *proportion* value is 100, the percentage of diabetic and non-diabetic consumers showing such behavior are equal. As an illustration, Figures 3 & 4 demonstrate two categories with different levels of predictor variables we initially selected for model-building.

MODEL-BUILDING

We split the entire data sample of close to 80,000 consumers into two sets. The first set (60% of this data sample) was used to train the model and the rest (40% of the data sample) was used to validate our model. An independent sample of 80,000 consumers was selected from the state of Missouri for out-of-sample testing purposes. As a consequence, the use of independent data to fit and test the model would demonstrate the generalizability of the model for the use of predicting outcomes for future subjects [3, 4]. The target value in the model has two possible outcomes: the person is (1) diabetic or (0) not diabetic.

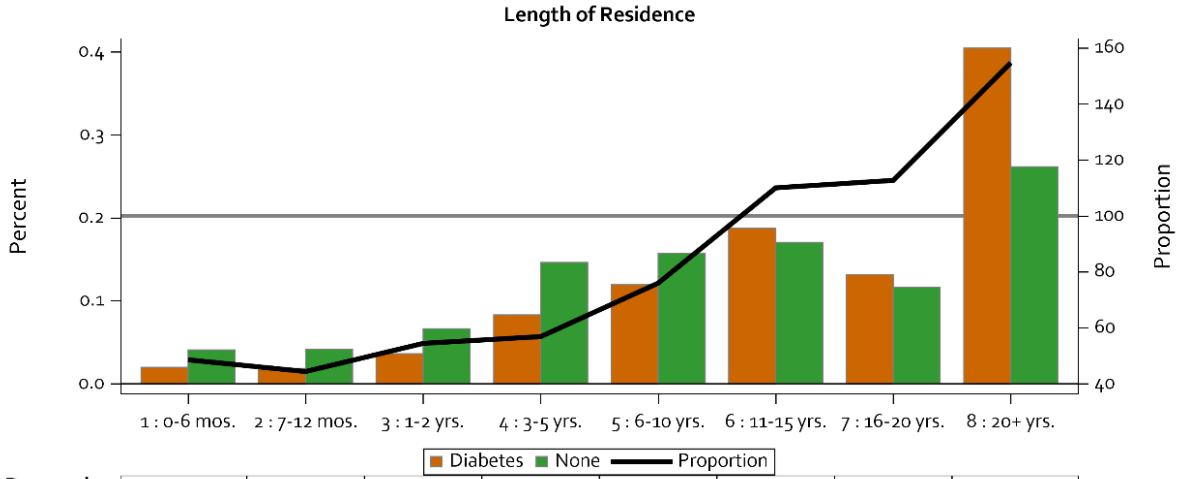
A: Variable Reduction:

We built a classification model (logistic regression) using SAS® Enterprise Miner, focused on demonstrating maximum discrimination between consumers with and without diabetes. Starting with over 1000 predictor variables, we eventually settled on the top 16 predictors for the model to maximize its predictability while keeping overfitting and multi-collinearity effects to the minimum. The predictor variable reduction method used in our analysis is summarized in Table-2.



Proportion	77.272	95.873	102.04	136.66	227.77
Diabetes	0.342	0.2047	0.1772	0.1676	0.1086
None	0.4426	0.2135	0.1736	0.1226	0.0477

Figure 3. Comparison of financial health of diabetic and non-diabetic consumers. The proportion curve shows individuals with below average and poor financial health are likely to be diabetic. In these cases, the proportion values are significantly different from 100 and meet our condition to be included in the modeling process.



Proportion	48.526	44.342	54.408	56.823	75.954	110.02	112.68	154.73
Diabetes	0.0198	0.0184	0.0361	0.083	0.1193	0.1873	0.1312	0.4046
None	0.0407	0.0415	0.0663	0.1461	0.1571	0.1702	0.1165	0.2615

Figure 4. Comparison of length of residence of diabetic and non-diabetic consumers. The proportion curve shows individuals with 11 years or more length of residence are likely to be diabetic while those with less than 10 years are not. In these cases, the proportion values are significantly different from 100 and meet our condition to be included in the modeling process.

Input variable selection method for target value (Diabetes: Yes or No)	Retained Dimension of Predictors
Chi-Squared test (based on correlation of binary variables with the target)	500 predictors
T-test (based on correlation of multivalued variables with the target)	300 predictors
Select variables with Information Value (IV) between 0.03 and 0.5, to rank the order of variables in terms of their predictive power	150 predictors
Perform variable clustering and select best variables based on *min*(1-R^2 ratio) from each cluster Drop binary variables with less than 10% occupancy Perform level clustering of categorical variables and merge levels which satisfies a Chi2 cut off with respect to the target	50 predictors

Table-2: Initial variable selection method used in the analysis to reduce dimensionality of predictors in the model

B: Logistic Regression

Consider that a collection of *p* independent variables that will cause an incidence of diabetes is denoted by the vector:

$$\mathbf{x} = (x_1, x_2, x_3, x_4, \ldots., x_p).$$

The conditional probability that the event (i.e. the person is diabetic) is observed can be denoted by:

$$p(y = 1|x) = \pi(x).$$

The logit of the multivariate logistic regression is given by the following expression, which is a standard in statistics literature [5, 6]:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots. + \beta_p x_p$$

Where g(x) = response to chronic condition (1 = diabetes, 0 = no diabetes), β_0 = intercept, β_1 = slope for independent (predictor) variable, x_1 = independent (predictor) variable1, etc.

Given the j^{th} independent variable has k_j levels, k_j – 1 dummy variables will be needed. We denote these variables by D_{jl} and their coefficients are denoted by β_{jl}, where $l = 1, 2, 3, 4, \ldots, k_j - 1$.

Hence the logit for the model with *p* variables and the j^{th} variable being discrete would be:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \ldots. + \beta_p x_p$$

The logistic regression is hence expressed by:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

We performed a step-wise logistic regression with the Schwartz Bayesian Criterion[7] to identify the top significant predictors by keeping the variables with p-value < 0.01, assuming the errors to be normally

distributed. Categorical variables, with a subset of levels found significant in the fit, are replaced by dummy variables for each significant level. We used the Lift and Cumulative Captured Response statistics to reject one of the predictors in any pair with more than 40% correlation. The same approach was repeated iteratively with the least significant variables until 16 predictor variables remained in the fit. The likelihood ratio for the *test for global null hypothesis* for this model was found to be 29337.45 leading to a p-value < 0.001.

MODEL PERFORMANCE

To assess model performance, we translated the model scores into deciles and compared the Cumulative Captured Response and Lift distributions for training and validation samples. At given decile, "lift" is defined as the ratio of predicted rate over average rate of incidence of diabetes in the sample. As shown in Figure-3, the highest decile has a lift of 2.4. That means compared to a random model capture of 10% of the total consumers our model is able to identify a maximum of 24% of the consumers who are diabetic. For a random model, the lift stays flat at 1, shown as a dashed line in the Figure-3.

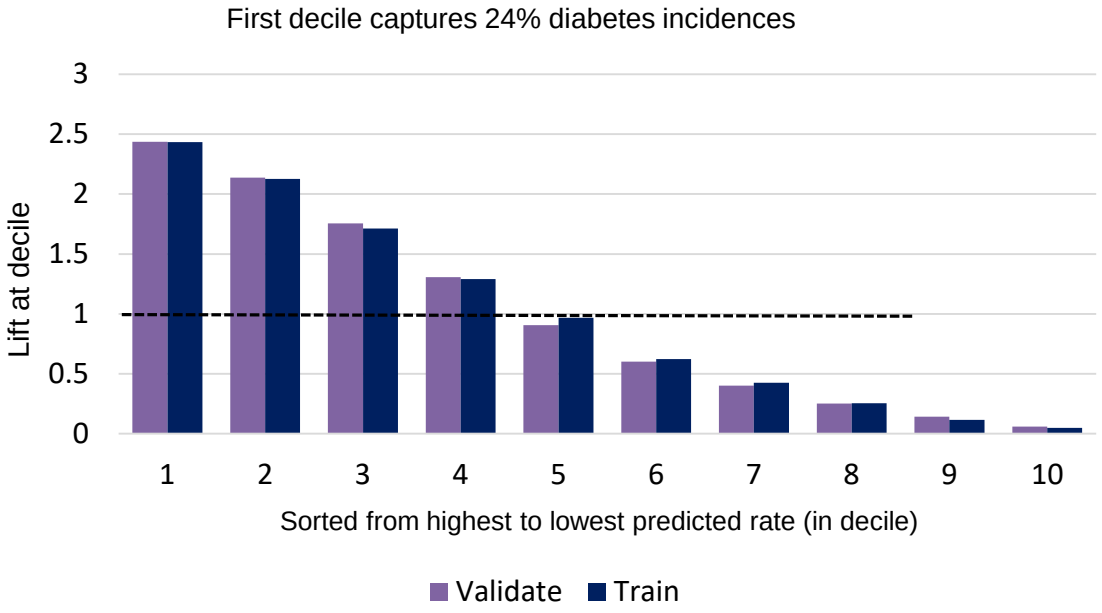


Figure-3: Observed Lift for the model is equal to 2.4. The top 10% individuals based on the predictive model show 24% diabetes incidences. Both training and validation data show very similar results in the five deciles.

Under the assumption of approximately 35 million people per decile (that would make the total US population to be roughly 350 million), the first decile of the model would correctly predict approximately 5 million (= 0.14 X 35 million) more people to have incidences of diabetes compared to a random model. Clearly the model does well to enhance events in the top five deciles, which are of most importance for targeting.

The Cumulative Captured Response is the rate of predicted diabetes among the consumers in a cumulative decile. For example, as shown in Figure-4, the cumulative diabetes prediction rate for consumers in top five deciles is 85%. In other words, the top five deciles (approximately 175 million people of the U.S.) of the model would correctly predict 85% of diabetic patients in the U.S. The 45-degree line on the graph represents a random model (50-50% chance of getting a right response as in a

coin toss). This further demonstrates the reliability of our model to predict diabetic incidence for a population, given its demography and lifestyle information.

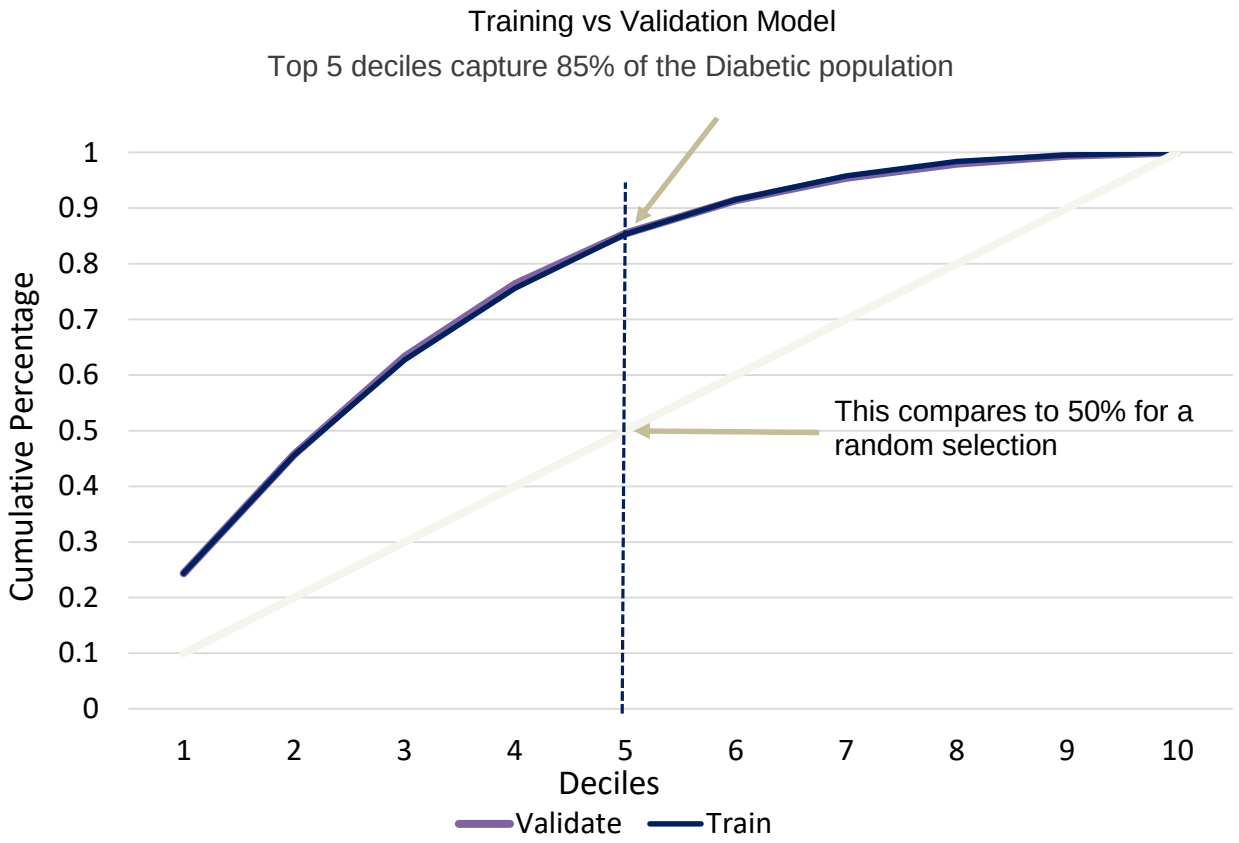


Figure-4: Cumulative gain chart shows that top 5 deciles capture about 85% of the diabetic population in the sample. This shows our model has practical predictive power. The gain chart for both Training and Validation data show very similar results.

For comparison, we have trained predictive models based on a boosted decision tree and a Neural Network on the same data sample. The former was found to perform marginally worse than the logistic regression model, while the latter was found to be either similar or slightly better across different deciles. For this study, we decided to adopt the logistic regression model for its simpler interpretation of relationships of the predictor variables to the target.

Predictor Variables for Diabetes Incidence:

We have identified 16 significant predictor variables, as shown in Table-3, that would predict diabetes for a person given his/her demography and lifestyle information (parameters) as indicated. The Maximum Likelihood Estimate table shows that all 16 explanatory parameters are significant in the model. Table-4 summarizes our interpretation on those predictor variables ranked from most significant to least significant. For example, according to our model, diet conscious consumers have the strongest association with diabetes. This could be a result of post diagnostic behavior. The model seems to have a common thread in the predictor variables that one could relate to “stress”. However, it is important to remember that stress can originate from different sources which are not necessarily correlated with one other.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-0.8057	0.0321	628.44	<.0001		0.447
Revolving Credit Card Users	1	-0.3255	0.0204	253.37	<.0001	-0.1794	0.722
Arts Events Patrons	1	0.4395	0.0185	567.09	<.0001	0.242	1.552
Diet Conscious Consumers	1	-1.1382	0.0199	3260.12	<.0001	-0.63	0.32
Entertainment Readers	1	0.5566	0.0179	965.35	<.0001	0.3062	1.745
Financial Institution Consumers	1	0.4008	0.0157	651.88	<.0001	0.221	1.493
Gospel Music Lovers	1	-0.2721	0.0173	245.97	<.0001	-0.1499	0.762
Interest Checking Account Holders	1	0.4253	0.0161	694.14	<.0001	0.2346	1.53
Mortgage Refinancers	1	-0.3379	0.0149	514.45	<.0001	-0.1865	0.713
Fast Food Restaurant Users	1	-0.6699	0.0189	1254.88	<.0001	-0.3691	0.512
Rewards Card Users	1	0.7709	0.0187	1691.54	<.0001	0.4249	2.162
Senior Caregivers	1	-0.763	0.0156	2387.89	<.0001	-0.4215	0.466
Bulk Item Shoppers	1	-0.5284	0.016	1095.9	<.0001	-0.2913	0.59
Vacation Spenders	1	0.4099	0.0209	384.73	<.0001	0.2258	1.507
Womens Plus Size Clothing Buyers	1	-0.1949	0.0182	114.34	<.0001	-0.1075	0.823
Net Worth Flag	1	-0.3493	0.0273	163.92	<.0001		0.705
Home Value	1	0.3243	0.0186	303.95	<.0001	0.1821	1.383

Table-3: Maximum likelihood estimates for the predictor parameters. We used Wald Chi-squared values to rank these predictors in the order of importance for predicting the incidence of diabetes.

Rank of Ordered Predictor Variables in the Final Model	Relationship with Response	Possible Explanation
Diet conscious consumers	Positive	More likely to become diabetic because of family history or simply showing post diagnostic behavior
Senior caregivers	Positive	More likely; high stress lifestyle
Rewards card users	Negative	Less likely; stress free lifestyle
Fast food restaurant users	Positive	More likely; unhealthy food habits
Bulk item shoppers	Positive	More likely; large volume consumer
Entertainment readers	Negative	Less likely; stress free lifestyle
Financial institution consumers	Negative	Less likely; affluent
Interest checking account users	Negative	Less likely; influenced by multiple factors
Arts events patrons	Negative	Less likely; affluent hobby
Mortgage refinances	Positive	More likely; demanding lifestyle
Vacation spenders	Negative	Less likely; stress free life
Revolving credit card users	Positive	More likely from low income group
Net Worth flag	Negative	Less likely to impact people with economic stability
Gospel music lovers	Positive	More likely; influenced by multiple factors
Home value	Positive	More likely; influenced by multiple factors
Women's plus size clothing buyers	Positive	More likely; weight factor

Table-4: The final dependent variables, ranked in terms of their level of significance in the model to predict diabetes condition. The possible explanations for each variable to show up in our predictive model is purely our interpretation without any help from experts in healthcare.

MODEL EVALUATION

In order to test the performance of the model classification, we scored an independent data set (from the state of Missouri) for which the true values (incidence of diabetes) are known. As noted earlier it is comprised of the same proportion of signal and background records as the modeling sample. The scores were translated to deciles following the same procedure as before. Table-5 summarizes the resulting Confusion Matrix.

Specificity is the ability of the model to correctly identify the individuals without diabetes (true negative rate). We compute that value to be: 5323 / (37140+5323) =12.53%.

Sensitivity is the ability of the model to correctly identify the individuals with diabetes (true positive rate). We estimate that value to be: 13819 / (13819+7264) = 65.54%.

Accuracy is an indicator of the overall effectiveness of the model across the entire data and we estimate that value to be: (13819+37140) / (44404+19142) = 80.02%.

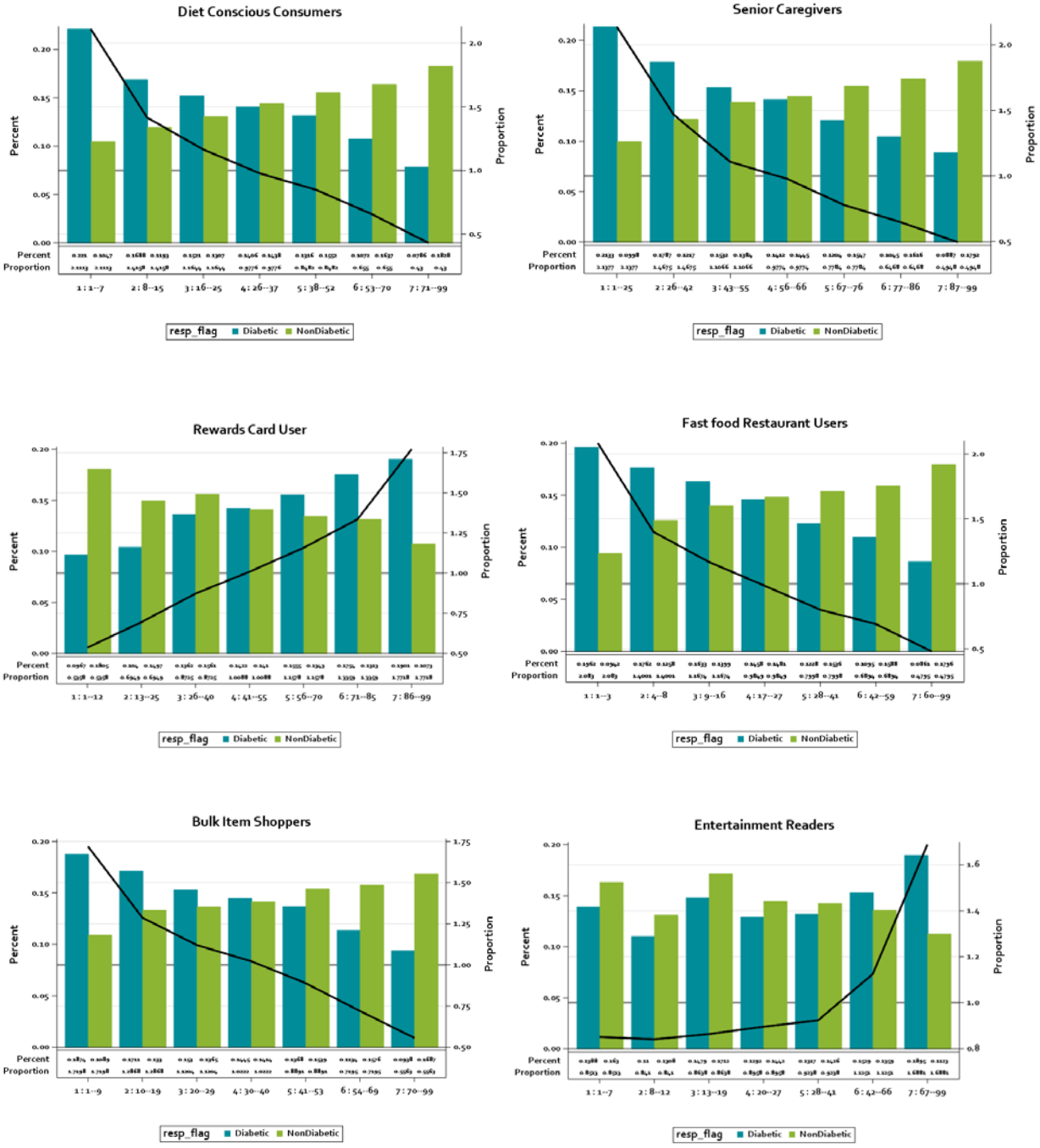
		PREDICTED TARGET		
		No Diabetes (Target = 0)	**Diabetes (Target=1)**	**Total**
Decile	**No Diabetes (Target = 0)**			
1			761	761
2			2108	2108
3			3174	3174
4		2892	1221	4113
5		4852		4852
6		5326		5326
7		5709		5709
8		5926		5926
9		6160		6160
10		6275		6275
Total		**37140**	**7264**	**44404**
	Diabetes (Target=1)			
1			5594	5594
2			4247	4247
3			3178	3178
4		1443	800	2243
5		1503		1503
6		1029		1029
7		645		645
8		429		429
9		195		195
10		79		79
Total		**5323**	**13819**	**19142**
		Specificity	**Sensitivity**	**Accuracy**
		=12.53%	**=65.54%**	**=80.02%**

ACTUAL TARGET

Table-5: Confusion Matrix with actual target (0/1) and predicted target (0/1) divided into deciles in rows and columns respectively.

DISCUSSION

The predictor variables in the model indicate some interesting insights into understanding chronic diabetes. The top 10 predictor variables were further visually explored with their levels (1 being most likely, 99 being least likely the person belongs to the specific behavioral category) grouped into seven bins, as shown in Figure-5.



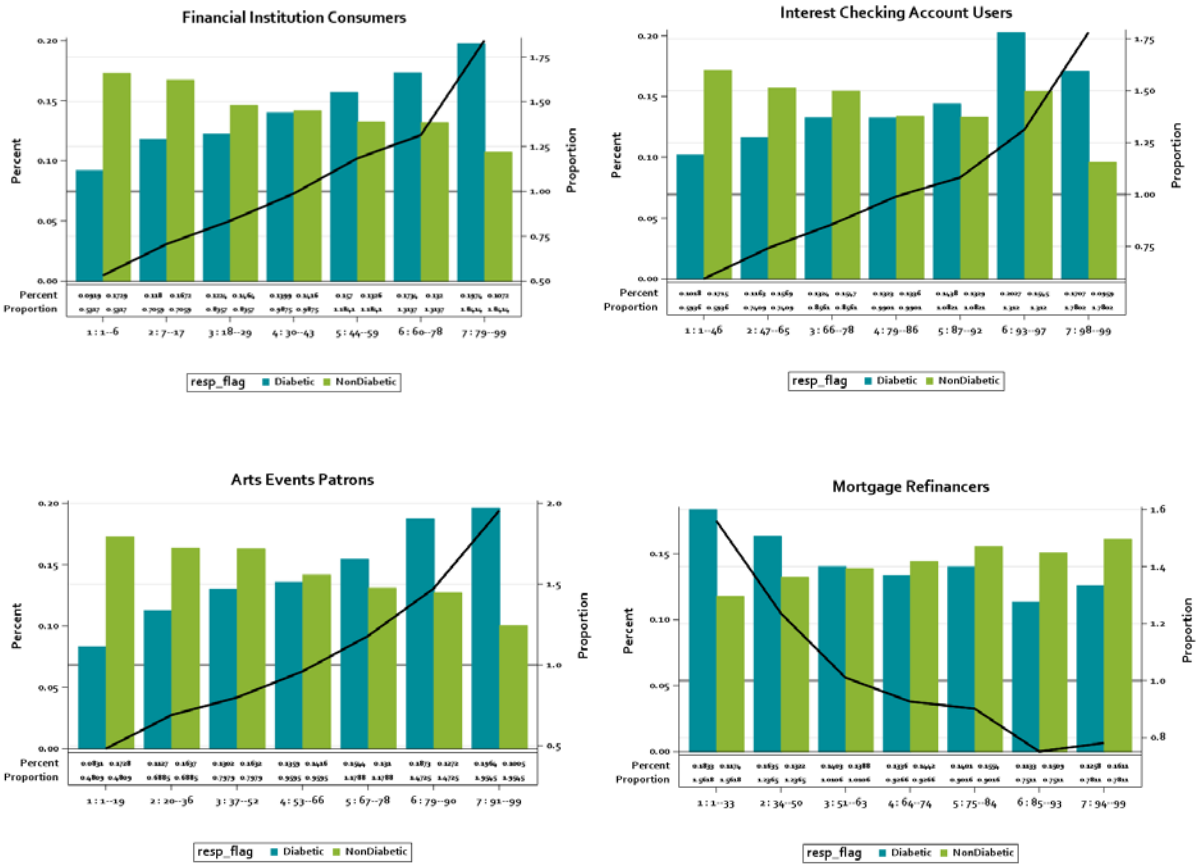


Figure-5: Charts of model variables comparing diabetic and non-diabetic consumers. The monotonically varying proportion curves (in black) confirm their ability to discriminate between the two classes.

The curve is the proportion of percentages of counts in each graph. Their monotonic variation further confirms the classification ability of our model.

Interestingly enough, most of these predictor variables are related to an individual's lifestyle, behavioral condition, and attitude towards life in general. Medical research has shown that stress, meals, and overall agility have significant correlations with the incidence of diabetes. For the first time, we have quantified these associations to predict diabetes from available consumer data that is mostly used for marketing research.

Although our model is tested and validated to have significant predictive power for classifying diabetes, we are not able to verify its accuracy on real subjects in the absence of accompanying clinical diagnostic information, and any similar studies to compare with. However, we believe our analysis opens up a new approach to study chronic diseases that one can utilize to compliment related clinical research.

CONCLUSION

Data mining techniques were used to analyze behavioral and lifestyle data and predict diabetes in individuals. Based on predictive modeling, we have identified 16 independent demography and lifestyle related variables that could be used to predict a person's propensity to become diabetic. Most of the identified predictor variables are based on consumers' socio-economic status, their social behavior and related action causing a level of stress.

We are motivated by the fact that most of these discovered variables are proactively manageable to avoid diabetes and we hope one can take these attributes to develop a technology application to engage people against such behaviors and lifestyles. This could potentially have an impact on people's lives and result in overall cost savings. However, the readers may be cautioned that initial assumptions about the data are critical to the applicability of the model.

We have demonstrated that consumer market research data can be used to complement clinical research to fight chronic diseases. We are encouraged to further expand our study to predict other prevalent chronic diseases, such as hypertension and high cholesterol.

REFERENCES

1. National Center for Chronic Disease Prevention and Health Promotion, National Diabetes Statistics Report, 2017, Estimates of Diabetes and Its Burden in the United States.
2. Division of Heart Disease and Stroke Prevention, https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_bloodpressure.htm
3. F. R. Harrell, K.L Lee, D. B. Mark, (1996), Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, "Statistics in Medicine", 15, pp. 361-387.
4. D. W. Hosmer, S. Lemeshow (1980), A goodness of fit test for the multiple logistic regression model, "Communications in Statistics", A10, pp. 1043-1069.
5. Aldrich, J. H. and Nelson, F. D. (1984) Linear, Probability, Logit, and Probit Models. London: Sage.
6. Hosmer D. W. and Lemeshow, S. (1989) Applied Logistic Regression. New York: Wiley.
7. Shtatland, Ernest & Kleinman, Ken & M Cain, Emily. (2003), Stepwise Methods in Using SAS® PROC LOGISTIC and SAS® Enterprise Miner for Prediction. SUGI ' 28 Proceedings: 2003.

ACKNOWLEDGMENTS

We thank the SAS Global Forum 2018 Conference Committee for providing a platform to present our results. We would like to express our gratitude to Prof. Goutam Chakraborty, Spears School of Business, Oklahoma State University, for his encouragement in the process. Dr. Asish Satpathy benefited from the support offered in part by the UCR's Unit 18 Professional Development Fund.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dr. Asish Satpathy
Olmsted Hall #2333
900 University Avenue
Riverside, CA 92527
asish.satpathy@ucr.edu

Dr. Asish Satpathy is a prolific entrepreneur, educator, and researcher who brings nearly 25 years of demonstrated success in (1) pursuing academic research; (2) managing mission critical particle detector instrumentation and data analysis projects in Europe (CERN), Japan (KEK), and the USA (Stanford University and University of California at Riverside); and (3) solving real world problems using innovative solutions that he created for his technology startup. His expertise in technology product innovation using Geographic Information System and passion to foster entrepreneurial spirit and creativity have influenced a significant number of business graduates and undergraduates from University of California at Riverside, where he serves as a faculty for the School of Business. His current research interests include application of novel statistical and visualization techniques to interpret data arising from public health and education studies.