Well-thy : Improve Health and Reduce Health Care Costs

Riken Shah, Ankitkumar Jain, Carolyn Thompson, James Henderson, Harry Aneja

Abstract

Healthcare costs are one of the primary attributes that impact virtually everyone. The goal of our system is to analyze habits and attributes of users to produce recommendations to improve their health. These recommendations reduce health care costs and show the user how much they could save by making the suggested habit changes. There are current systems that give healthy habit suggestions and financial recommendations separately, but no current application can quantitatively define the health attributes with actual dollar value of savings. Wellthy fully integrates health and finances to help users save money while they become healthier.

I. Introduction

U.S. health care spending reached $3.3 trillion or $10,348 per person in 2016, and health spending accounts for 17.9 percent of the nation's Gross Domestic Product.[1] National health care spending is projected to increase up until 2026 at least.[1] Health care costs per individual can be influenced by health status, hobbies, occupations, and habits. In this paper, we propose the Wellthy model that analyzes the attributes and habits of the user to improve their health and reduce associated health care costs.

II. Critical Thinking Steps

Critical thinking consists of various deliberative processes aimed at making informed decisions about actions and beliefs. The following critical thinking steps outline major project considerations:

- **Recognize problem:** Developing a model to analyze the attributes and habits of the user to improve their health and reduce the associated health care costs.
- **Gather relevant information:** Attributes that affect health care costs, potential data sets, less intuitive factors that affect insurance rates, similar existing systems.
- **Recognize unstated assumptions and values:** Initial assumptions about the weights of attributes.
- **Lessen ambiguity:** Specifying problem statement, relevant health attributes and proposed implementation.

- **Interpret data:** Two datasets with quantitative healthcare attributes were merged to calculate feature importance.
- **Recognize logical relationships between propositions:** Gradient boosting was used to find feature relevance weights, which are used to compute user health scores.
- **Draw/Test conclusions and generalizations:** Proof of concept demonstrating how health parameters affect healthcare costs

III. Motivation

The problem of creating an insurance ratesetter was assigned to our team as part of CSC 495/591, Data Driven Decision Making. To lessen ambiguity of the project goal, the problem was redefined as a model that takes user attributes and creates recommendations to reduce health care costs. Research conducted during the project demonstrated that habits affect health care costs, and we wanted the Wellthy model to address this by providing lifestyle change suggestions paired with monetary savings.

IV. Current Systems

There are current systems that provide financial recommendations or health recommendations but do not integrate the two.

- **Exeq[2]:** This company provides financial recommendations based on your spending habits. Our app will provide recommendations for changing lifestyle as well as spending habits without requiring the user to put in their credit card.
- **Health[3]:** This app gives a user recommendations to be more healthy by walking, running or doing some kind of exercise. Even though it has the recommendations, it can't be directly linked to decrease insurance rates.
- **Progressive Life Insurance[4]:** The customer isn't told what attributes affect the insurance rate when getting a quote or how to lower the quote. Our application will suggest habit changes to lower rates and will show how much the user saves.
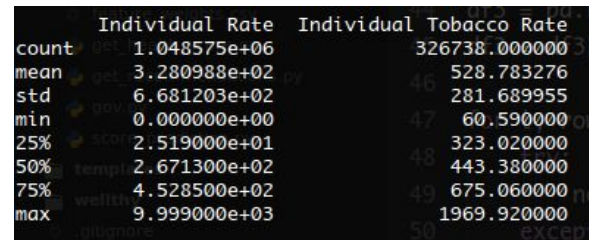
V. Data Sources

We looked at primarily two datasets which helped us quantitatively define the impacts of various healthcare attributes like smoking, drinking, BMI, etc. The major challenge with these datasets was to combine the quantitative information of non-overlapping attributes in these datasets to form a single dataset which could be used for model generation. We used a reverse-mapping technique for the fusion of this datasets and we were able to generate considerably good results with it.

The first dataset was obtained from Prudential life insurance assessment challenge posted on Kaggle[5] . The various attributes that were present in this dataset are as follows : Employment Info, Insured Info, Insurance History, Medical History, Family History, Age, Height, Weight, BMI. These attributes like age, height and weight were most important in our model since other attributes were normalized in 0-1 with no information about the denormalization procedure, preventing us from reconstructing the normalization module.
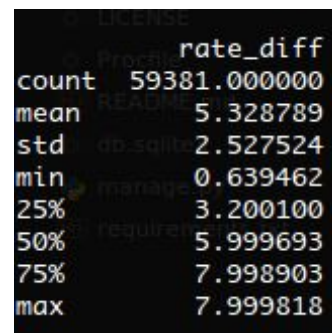
The second dataset was obtained from healthcare.gov[6]. The attributes of these dataset are Business Year, Issuer ID, Source Name, Version Number, Import Date, Tax Identification Number, Rate Effective Date, Rate Expiration Date, Plan ID, Rating Area ID, Tobacco, Age, Individual Rate, Individual Tobacco Rate, Couple, Primary Subscriber and One Dependent, Primary Subscriber and Two Dependents, Primary Subscriber and Three or More Dependents, Couple and One Dependent, Couple and Two Dependents, Couple and Three or More Dependents.

The major task was to combine these two datasets in order to find relations between attributes and their related effects on healthcare costs. We used two attributes from one dataset namely individual rate and indiv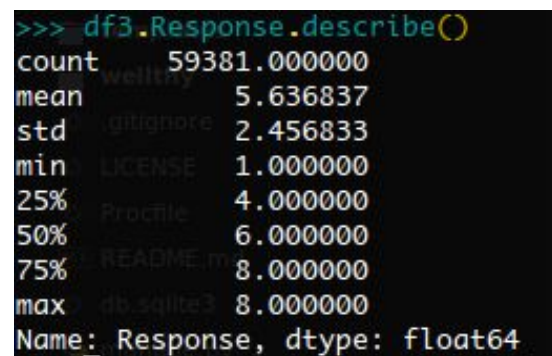idual tobacco rate to map the response parameter of the other dataset. The underlying assumption here was that higher difference between individual rate and individual tobacco rate would correspond to higher healthcare costs. The summary of this process can be illustrated by following images.

	Individual Rate	Individual Tobacco Rate
count	1.048575e+06	326738.000000
mean	3.280988e+02	528.783276
std	6.681203e+02	281.689955
min	0.000000e+00	60.590000
25%	2.519000e+01	323.020000
50%	2.671300e+02	443.380000
75%	4.528500e+02	675.060000
max	9.999000e+03	1969.920000

	rate_diff
count	59381.000000
mean	5.328789
std	2.527524
min	0.639462
25%	3.200100
50%	5.999693
75%	7.998903
max	7.999818

```
>>> df3.Response.describe()
count    59381.000000
mean         5.636837
std          2.456833
min          1.000000
25%          4.000000
50%          6.000000
75%          8.000000
max          8.000000
Name: Response, dtype: float64
```

As illustrated in these images, we mapped the two datasets by Response factor (which represented a risk associated) and normalized rate difference between users who consumed tobacco and those who didn't. These mapping preserved the correlation between attributes and also gave us a

combined and consolidated dataset which would have all the attributes that we aimed to include in the system.

Additional data sources were used to justify weights of health attributes. The first describes the health-care costs smokers pay in their lifetime and over five years as well as how much they would save by quitting.[9] The second describes a list of common health conditions that affect life insurance costs.[10] Another source describes the incremental health effects and health costs of obesity.[11] Lastly, an article analyzing the health costs of drinking was referenced.[12]

VI. Implementation

Technology stack :
- **Python** : We used python in the backend to build the model and other functionalities. The recommendation engine as well as the prediction model is developed in Python.
- **Pandas :** We used pandas to load the data and perform different operations on the data. The slicing and other data manipulation methods of pandas were very useful in preprocessing steps.
- **Scikit learn :** We used this library to make use of in-built machine learning packages in python.
- **Django :** We used django to build the UI and take user input. It

provided a nice MVC architecture incorporating separation of concerns as well as faster development.
- **Github :** We used github to do version control and collaborate amongst each other.

Calculate Feature Importance : We merged the datasets we obtained from healthcare.gov and prudential health by normalizing the difference in healthcare cost of people consuming tobacco and not consuming tobacco and mapping it with the 'response' column in healthcare.gov data. Then to find the weight of each feature by which it influences the target we used machine learning techniques like Linear Regression and Gradient Boosting Regressor. Gradient Boosting for regression builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

To select the model we calculated the R-squared score of all the models. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. Out of all models we built, Gradient Boosting had the highest R-squared score hence we used it to find the feature relevance weights. These weights were used to

find the total health-score of a user. The model also provides if the attributes affect the health score positively or negatively. Following is the features, their weights obtained and effect -

Feature	Weight	Effect
age	0.1	negative
bmi	0.2	negative
ailments	0.2	negative
tobacco	0.1	negative
smoke	0.1	negative
drink	0.1	negative
exercise	0.05	positive
travel_time	0.05	negative
sleep_time	0.05	positive
job_type	0.05	negative

Build user profile :
This feature takes all the input parameters from the user and it gets stored in the sqlite database. The attributes are highly configurable using the class based views that django provides. The range of values are dynamically selected from model file which corresponds to the feature weights that we obtained from the machine learning model. The highly configurable nature of the profile makes updates very easy hence incorporating different weights of parameters for experimentation. Also the layout uses responsive bootstrap design taking into consideration cross-platform compatibility of the system.

Calculate Health Score and Cost Savings : Once the user profile is built we do statistical and textual analysis on different attributes dynamically and generate the health score and the amount in dollars the user can save on his total health costs if he abides by the recommendations provided by the system. Steps to calculate health score are as follows :

- **Preprocessing :** Here we do text analysis on the Past Ailments and Job Type input against the bag of words to find out if the user has any past ailments or if he/she works in a type of job that can adversely affect the health and incur high health-care costs. We calculate the BMI based on the height and weight input and see if it is in the healthy BMI range of 18.5 - 24.9 for adults[7] Similarly, we find out if the person belongs to the healthy age of 18-45 as health-care premiums increase greatly outside that range[8]. Finally, we encode the inputs for 'Drink', 'Tobacco', 'Smoke' as 1/0 based on true/false, respectively, provided by the user.

- **Formula to compute health score :** We came up with a robust, flexible and scalable formula

calculate the health score of a user based on his profile. It is as follows :

If effect is negative score is :

$$\sum_{i=1}^{n}[M - (M * val_i/opts_i)] * W$$

If effect is positive score is :

$$\sum_{i=1}^{n}[M - (M * (opts_i - val_i)/opts_i)] * W$$

where,

n = total number of attributes

M = maximum health score

val = value of the attribute

opts = number of segments in the attribute

W = feature weight

For example:

Travel Time per week:

< 5 hours → 0

5 - 10 hours → 1

> 10 hours → 2

If max healthscore = 1000 and weight = 0.1 and user selects 5 - 10 hours (1) then,

Healthscore = [1000 - 1000 *(1/3)] * 0.1

Healthscore for Travel Time = 66.67

Similarly, we calculate healthscores for each attribute and add them.

- **Formula to compute cost savings :** Cost that can be saved is directly proportional to the difference in the maximum health score and the current health score and also depends on the current total health care

expenditure of the user. The formula we used to compute the savings is as follows :

$$savings = points * totEx/M$$

where,

points = healthscore that can be increased

M = maximum health score

totEx = total health care expense of the user

For Example:

If healthscore = 900 then scope for increase = 100

Total expenditure per year = $500

then,

Savings = 100*500/1000 = $50

The following image shows UI representation of the user profile.

Following is your health profile

Attribute	Value
Age	21
Height	64.0
Weight	120.0
Ailments	
Health Care Costs (per year)	4300.0
Tobacco	False
Smoke	False
Drink	True
Exercise	<6 hours/week
Travel	<5 hours/week
Sleep	6-8 hours/day
Job Type	researcher and writer

Generating Recommendations

Once the health score and savings has been calculated, the system will generate recommendations for the user based on their profile's lifestyle factors and weights previously discussed. The application will provide the user with a total recommendation that includes all the lifestyle factors they can improve. It will also provide the user with single recommendations for each lifestyle factor. The recommendations includes a healthscore improvement feature which calculates the point improvement if you follow the recommendations.

For recommendation we have used a templates and rule based approach where the final recommendation text is generated by combining individual templates dynamically based on attributes values. The final text is completely configurable and totally depends on user profile.

- **Calculating Healthscore Improvement:**
  I = M * weight / n
  I = point improvement
  M = maximum healthscore
  n = number of possible values for feature.

The following image is an example of a generated total recommendation.

Health Score (out of 1000)	Possible Savings (per year)
900.0	$430.0

Recommendations

If you increase your exercise to atleast 6 hours a week, increase the amount you sleep to above 8 hours a day, and stop drinking your healthscore will improve by 100.0 points.

The following image shows a breakdown of the total recommendation shown previously.

If you increase your exercise to atleast 6 hours a week your healthscore will improve by 33.33 points.

If you increase the amount you sleep to above 8 hours a day your healthscore will improve by 16.67 points.

If you stop drinking your healthscore will improve by 50.0 points.

VIII. Evaluation & Results

We have created a proof of concept that calculates a health score given a user profile, computes cost savings, and generates recommendations for lifestyle factors the user can improve. We conducted manual evaluations of the system in order to justify the correctness of the system and with enough data, the validations can be automated. Also, large user surveys could be done in order to rate the recommendations that the system generates. Overall, in the set of the attributes that the system takes as an input, the results that are generated demonstrates dynamicity and correctness of the system.

IX. Future Scope

Wellthy could connect to fitness apps and wearables like Fitbits to pull health data such as weight, height, amount of exercise and walking. This would speed up the questionnaire step. Wellthy could also become a personal health companion, which will involve the personification of the app and more personalized health recommendations over time. Wellthy can be converted into a social platform that allows users to share their health updates and give advice to other users.

XI. Conclusion

The Wellthy model analyzes habits and attributes of users to produce recommendations to improve their health. To create a product which can help users become more healthy and save money on health care costs; We looked at primarily two datasets which helped us quantitatively define the impacts of various healthcare attributes like smoking, drinking, BMI, etc. By working with this data we were able to create a health score scale and provide recommendations to the user based on their health score. Based on the health score and the data entered by the user we are able to calculate the health care cost savings as well.

The recommendations generated by the application help the user lead a healthy life and save some money on health care costs along the way. This application can be expanded significantly by incorporating health data from devices like apple watch, fitbit and other wearables. This would help us gain more dynamic data about the user and can help us give live recommendations based on the users daily activity.

References

1. NationalHealthAccountsHistorical. (2018, January 08). Retrieved April 12, 2018, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html
2. Exeq. (n.d.). Retrieved April 17, 2018, from https://exeq.com/
3. IOS - Health. (n.d.). Retrieved April 17, 2018, from https://www.apple.com/ios/health/
4. Life Insurance Quotes - See Life Rates Now. (n.d.). Retrieved April 17, 2018, from https://www.progressive.com/life
5. https://www.kaggle.com/c/prudential-life-insurance-assessment/data
6. Rate PUF - 2015. (n.d.). Retrieved from https://data.healthcare.gov/dataset/Rate-PUF-2015/r3nu-ebdw
7. Body Mass Index (BMI) for Adults https://www.webmd.com/a-to-z-guides/body-mass-index-bmi-for-adults
8. How Age Affects Health Insurance Costs https://www.valuepenguin.com/how-age-affects-health-insurance-costs
9. Boonn, A. (2014, December 18). HEALTH COSTS OF SMOKERS vs. FORMER SMOKERS vs. NON - SMOKERS AND RELATED SAVINGS FROM QUITTING. Retrieved April 26, 2018, from https://www.tobaccofreekids.org/assets/factsheets/0327.pdf
10. Life Insurance with Pre-Existing Conditions. (n.d.). Retrieved from https://www.aigdirect.com/about-life/planning-for-life-insurance/preexisting-condition-life-insurance
11. Fallah-Fini, S., Adam, A., Cheskin, L. J., Bartsch, S. M., & Lee, B. Y. (2017). The Additional Costs and Health Effects of a Patient Having Overweight or Obesity: A Computational Model. Obesity, 25(10), 1809-1815. doi:10.1002/oby.21965
12. Hunkeler, E. M., Hung, Y., Rice, D. P., Weisner, C., & Hu, T. (2001). Alcohol consumption patterns and health care costs in an HMO. Drug and Alcohol Dependence, 64(2), 181-190. doi:10.1016/s0376-8716(01)00119-3