Executive Summary

Xuan Li (xuanli1)

General health examinations (GHE) has been the standard practice for years. Many people believe that it will help for early detection of potential illness. Besides the relatively expensive service fee, however, there might exist other factors, i.e., the professionality of medical staff, or attractiveness of advertisement, that would hinder people in less developed countries from conducting the check-up regularly. Using a survey dataset queried from 2068 respondents in Vietnam, in this project the most important factors for regular check-ups are identified with a random forest classifier. Furthermore, to make the advertisement more effective to a broader audience, an analysis on the potential people that are more likely to go for regular check-up is also conducted. The general findings are summarized as follows.

Essentially, people are more satisfied with the professionality (*check-up quality*) of the medical staff, include their reliability, responsibility, empathy, etc. On the contrary, they usually give lower rating to the information (*check-up information*) they receive during check-ups, especially they don't believe the information are attractive or impressive enough. However, due to the similarity and ambiguity between the terms, it is highly suspicious that these ratings truly reflect the evaluations independently.

In addition, when making decisions for regular check-ups, common factors have been detected that have equivalent importance for people with/without symptoms of illness. The most important factors include age, BMI, job status and amount of time suitable for exercise. Also, while *check-up quality* are more crucial for people without symptom, for people with potential illness the quality of *check-up-information* have a higher weights. On the other hand, factors including whether the respondent or family is in good health, takes simple medical measurements regularly, or has little faith in the quality of medical service can be ignored due to their insignificant impacts.

Lastly, with a logistic regression model we have identified 40/369 people that don't get check-up regularly can be potentially persuaded with little effort. There is only a few proportion of people that would be extremely hard to change their mind. Although no unique pattern have been revealed from the people can be easily converted, they usually possess the features of having a higher degree, owning a health insurance, and working out often.

Introduction

General health examinations (GHE) is crucial not only for individuals to remain alert on potential illness, but also for medical institutions since it makes up of a great proportion of their annual incomes. Although the expense can be nontrivial, it might not be a big issue for most people in developed countries as it can be covered by family insurance or employee as part of the benefit. Along with other factors, however, it can be a major concerns that inhibits individuals to do the check-up regularly in less developed countries, which is of specific interests from researchers in Vietnamese Ministry of Health. To investigate the most important factors that motivate people to do GHE regularly, the researchers conducted a country-wise survey in 2016, collecting information from a variety of entities including schools, companies, government and households by conducting 15 minutes survey.

The dataset contains 2068 valid responses from the 2479 people that have been approached to in year 2016, which is downloaded from the link below[1]. Besides the unique id of survey respondent, the questionnaire covers 49 general queries, including objective information such as age, gender, height, and respondent's subjective opinions such as the respondent's rating on the quality of medical equipment (Tangibles), ability of examiner to perform medical services (Reliability) and how impressive the information they receive in check-ups (ImpressInfo). The two most important information in the dataset are the RecExam and RecPerExam, which are the time since the respondent last visited a doctor given any symptoms of a disease and without specific illness.

The researchers are curious about the public opinions on the professionality of the medical staff and how alluring the information provided during check-up. Besides, with limited advertising budgets at hand, they want to emphasize the most important factors in the brochure, and make it target towards the people that can be easily converted for regular check-ups. Thus, in this project we would implement specific analysis and classifier, i.e., random forest and logistic regression, to provide insights to each of the concerns. Finally, the summary of discovery would be given in the conclusion.

Data Exploration and Diagnosis

Although all of the entries in the dataset are stated as 'valid', we have identified several suspicious inputs that need to be handled before getting into analysis. Firstly, there are 8 outliers in date that are marked as 20169828, which is clearly casued by misspelling. These entries are adjusted as 20160828, based on the conjecture that 0 should be the 'nearest possible neighbor' on keyboard for 9. Another observation is that there are a small proportion of entries that are marked as '2,5', '3,5' in the columns that should have recorded the rating from 1 to 5. As inferred from the BMI column, however, the comma here might work as decimal point. Thus, to make these rating consistent we just floored down the value to its nearest integer.

On the other hand, we have also found some potential factors that would introduce bias. One such factor is the Age of the respondent. As showed in Figure 1, most of the people that participated the survey are younger people aging from 20 to 30, which makes the distribution highly skewed towards this range. The log transformation, however, have resolved this issue to a certain degree. Thus, the log transformed version for Age would be used in this analysis. In addition, after adjusting the outliers in the date, it is observed that 99% of the surveys are conducted during late September and early October, which does not make this factor ideal as an covariate. For the reason above we will ignore this factor in the following sections.

[1] `http://rosmarus.refsmmat.com/datasets/datasets/vietnam-health/`

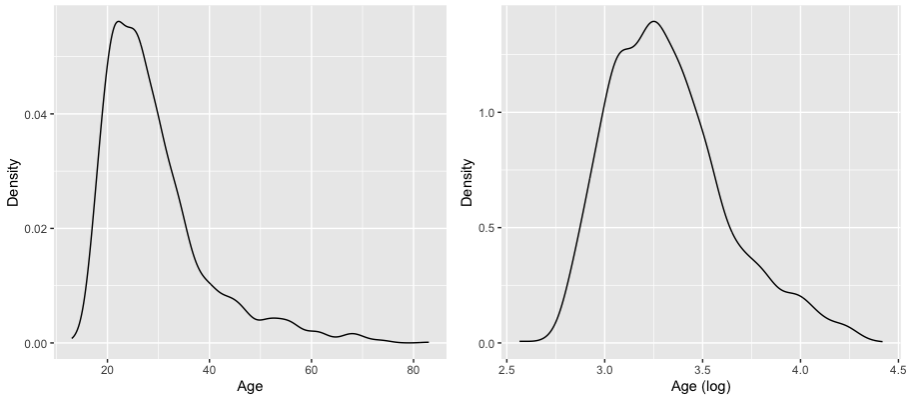

Figure 1: Less skewness of Age after log transformation

Method and results

Part I: outlook of the rating for information received at check-up, and professionality of the doctors

Firstly, we have noticed there are several questions that ask respondent's rating on the information provided in check-ups, including sufficiency (SuffInfo), AttractInfo (attractiveness), impressiveness (ImpressInfo), and popularity (PopularInfo). Besides, there is another set of rating related questions regarding on professional skills of the nurses and doctors, which are tangibility of equipments (Tangibles), reliability (Reliability), timeliness of service (Respon), knowledge of the doctors (Assurance), and thoughtfulness of the medical staff (Empathy). Above all, these subjective ratings can be crucial for individuals to make decision on whether to go for check-up regularly. On the other hand, they also represents the aspects that medical center can actually improve quickly without sophisticated equipment upgrade. To resolve the above concerns we first look into the histogram of the two group of factors, which is showed in Figure 2. We use the term of *check-up information* and *check-up quality* to represent the factors respectively.

From the histogram it is observed for both group most of the rating are centered around 3, which is normal since the respondents only have limited time (12-15 minutes) to response. Given the large number of questions they need to answer, simply filling with neutral values might be the best option they have in order to complete the interview in time. However, in general there are more low ratings in *check-up information* while in *check-up quality* more high ratings are given. A direct indication is that most of the people are satisfied with the expertise of doctors and service from nurses. Specifically, they are contented with the quality of medical equipment and personnel, while there is still room for medical staff to improve their thoughtfulness and sense of responsibility. However, the information provided at check-ups are somewhat problematic, where ratings ≤ 3 dominate the questionnaire. In particular, most people neither believe the information is attractive nor impressive, even though the material is slightly comprehensive as they might have expected.

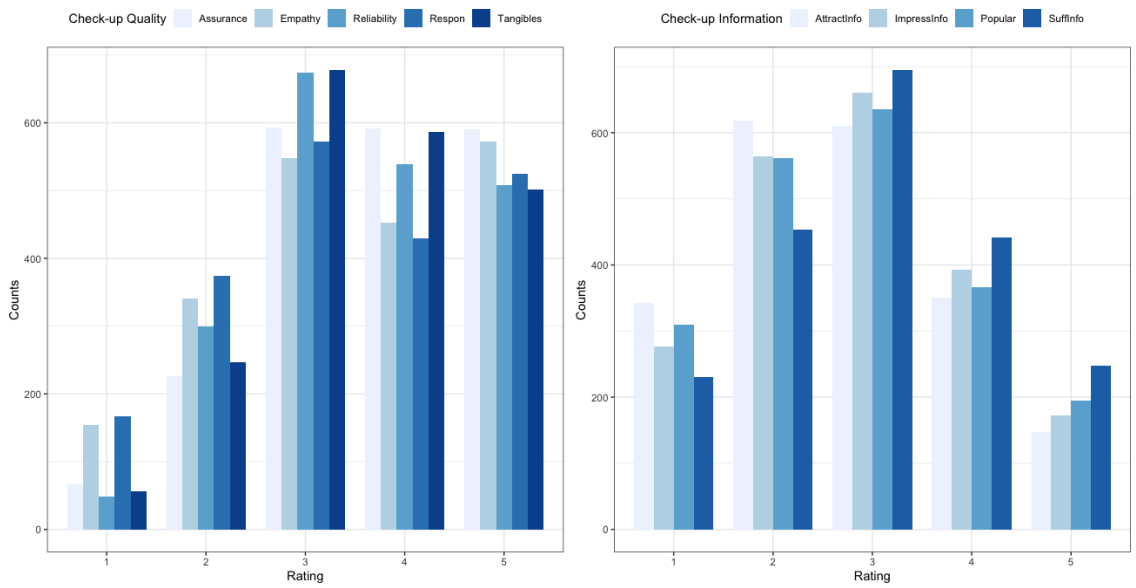

Figure 2: Most ratings are neutral: left-skewed *check-up information*, but right-skewed for *check-up quality*

However, from the histogram as well as the description of the questions, it seems that it is hard for the respondents to distinguish these terms well, which may ruin the value of these queries. In other words, these features might not satisfy the i.i.d. assumption for covariate that is required by most of the models. The fact is reflected in Figure 3, where relative high positive correlations have been detected for pairwise factors within *check-up information* as well as *check-up quality*. As expected, it is neither easy to tell the difference between reliability and assurance, nor between impressiveness and attractiveness of the information. As a result, however, the existing ambiguity in definition might make introduce some bias in the actual ratings.

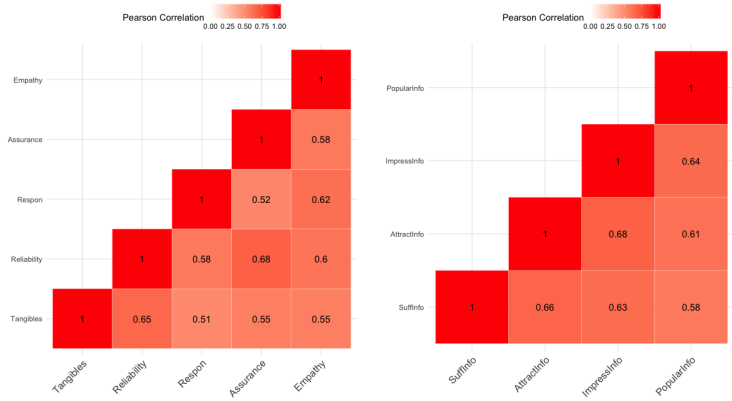

Figure 3: High correlations observed within *check-up quality* and *check-up information*

Part II: factors that motivate check-ups every twelve months

Considering that the medical institutions might only have limited budget for advertising, it would be wise for them to emphasize specific information in the brochure and make it target towards the

people who are more likely to go for check-ups regularly. Specifically, the factors that make people to go for check-up every twelve months need to be revealed and quantified to improve advertising strategies. To account for the cases that whether a person gets check-up with specific symptoms, RecExam and RecPerExam would be used as response variable individually. The entries labelled as 'less12' are regarded as one class (marked as 1), and the rest as the other class (marked as 0). Without loss of generality, we want to build a classifier that not only distinguishes the two class with certain guarantees, but also provides a measurement of the importance of all the covariates considered. As a consequence, random forest comes on top from the candidates, given its consistent and superior performance in industry applications. On the other hand, the importance of each features can be calculated and compared based on how much impurity it decreases, where the impurity can be regarded as a surrogate of the loss function that random forest aims to minimize. Also, coming along during the training stage, the out-of-bag error rate of random forest can be used as an approximation of the generalization error, which eases the pain of cross-validation. Thus, random forest classifiers would be implemented for this part of analysis.

We build two random forest classifiers using the default setting for the two response respectively. Expect date and Age_gr, all of the factors are used as covariates for both of the model. The confusion matrix of the out-of-bag error for the two cases are showed in Table 1. Without specific tuning the model can perform fairly well for both of the cases and an accuracy of around 0.75 can be expected. However, the model suffers from data imbalance issue for people with symptom case, where a great proportion of samples are predicted as false positive.

with symptom	pred 0	pred 1
true 0	334	361
true 1	174	1199

no illness	pred 0	pred 1
true 0	774	235
true 1	262	797

Table 1: Confusion matrix: accuracy of 0.74 and 0.76, respectively

Additionally, using impurity as the measure of importance, the ranking of critical factors are showed in Figure 4 and 5 respectively. As briefly mentioned above, the magnitude of impurity can be regarded as its contribution in terms of loss reduction. In other words, the model would select such factors as decision node when splitting data. Although the ranking may vary slight, the top 20 most important factors for both of the cases are similar. For example, Age, BMI, job status and amount of time suitable for exercise are the top 10 factors for regular check-ups with or without symptom. Also, all of the five *check-up quality* factors are within top 20 for the symptom case, while all of the four *check-up-information* ones hold for the case with no specific symptom. On the other hand, the most insignificant factors are also similar as well. For instance, whether the respondent or family is in good health (StabHthStt), takes simple medical measurements regularly (ExamTools), or has little faith in the quality of medical service (Lessbelqual) does not necessarily affect his/her frequency to do the check-up.

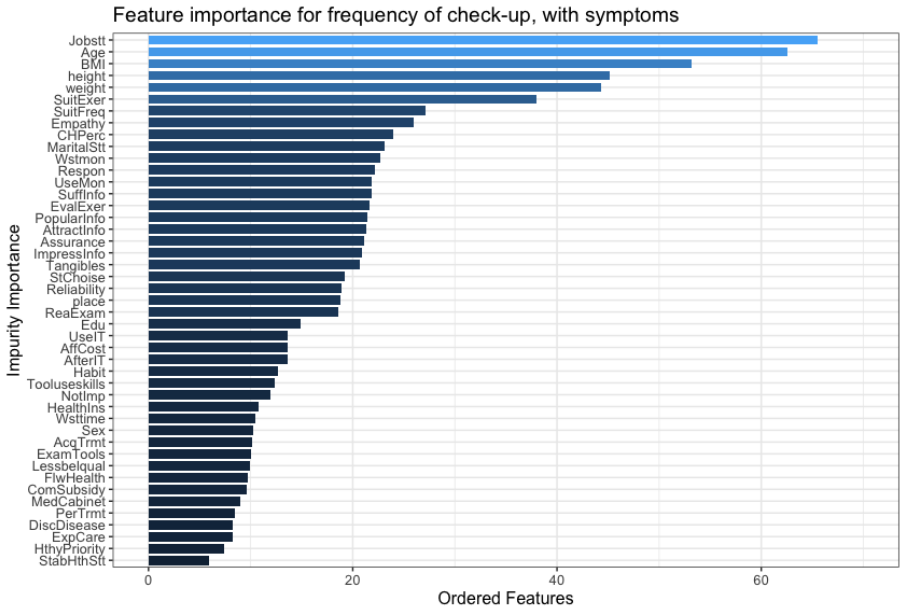

Figure 4: Top three factors for check-up with symptoms: job status; age; BMI

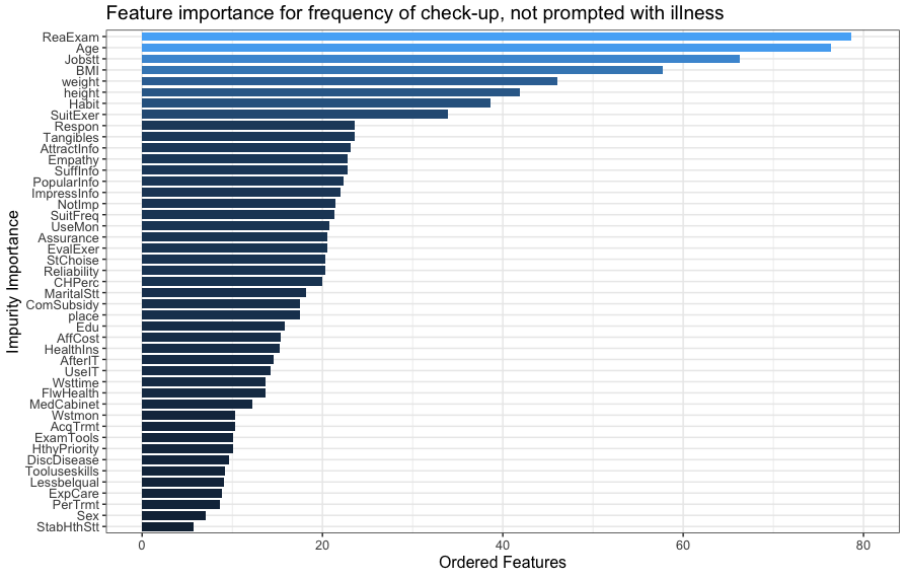

Figure 5: Top three factors for check-up without illness: reason for last check-up; age; job status

Part III: persuading people on the edge

As reflected in the original dataset, 66% of the respondents indicated that they went for check-up within 12 month for specific symptom of disease. While some of the remaining 34% might be stonehearted on whatever advertisement they are provided with, the rest can be persuaded to change their mind with little push. To identify the presence of patients with such characteristics, we need to come up with a classifier that can also generate a reference likelihood for these people to switch.

Logistic regression would be an ideal choice due to the above considerations. Basically, using the log odd inference showed in Equation 1, the probability of whether a person will go for check-up regularly (labelled as 1) can be calculated with the fitted coefficients β_i. Thus, those who don't get check-up regularly but are predicted with high $P(y = 1)$ while $P(y = 1) \leq 0.5$ can be regarded as the people can be easily persuaded.

$$\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \sum_i^n \beta_i x_i \tag{1}$$

Based on the above mechanism a logistic regression model is fitted using all possible factors to predict the response RecExam. date and Age_gr are excluded for similar reasons as stated earilier, People who get check-up within 12 month are marked as class 1, and all other cases are marked as class 0. Fitting the model with all entries, the significant covariates are reported in Table 2 below. In general, people who think check-up is a waste of money, not important, don't have the regular check-up habits are definitely less likely to go every twelve month. On the other hand, the odds is also dim for people who are unmarried and without stable job. Finally, it seems that the *check-up information* or *check-up quality* would not have significant impact for people with symptom of disease when making decisions. They might be prompted, however, if they feel the information is impressive.

covariate	**coefficient**	$\Pr(>\|z\|)$
Age	-0.029	0.001
Sex (male)	-0.397	0.045
Jobstt (student)	-1.318	0.001
Jobstt (unstable)	-0.946	0.049
MaritalStt (unmarried)	-0.431	0.010
HealthIns (yes)	0.446	0.002
Wstmon (yes)	-0.291	0.040
NotImp (yes)	-0.321	0.007
Habit (yes)	0.341	0.008
UseMon (later)	-0.350	0.012
UseMon (partly)	-0.479	0.002
Empathy (5)	0.779	0.008
StChoise (clinic)	0.277	0.043
StChoise (selfstudy)	0.447	0.004
CHPerc (good)	-0.366	0.038
SuitFreq (18m)	-0.841	0.01
SuitFreq (g18m)	-0.723	0.002
ImpressInfo (2)	0.472	0.04
SuitExer	-0.004	0.035
EvalExer (quitesuff)	0.551	0.0001

Table 2: Coefficient of significant covaraites

Suffering from data imbalance issue, from Table 3 it is observed that half of the people that don't get check-up regularly as identified as the opposite, resulting in an overall prediction accuracy of 0.76. Using ten fold cross validation, the generalization accuracy is about 0.82.

no illness	pred 0	pred 1
true 0	369	**326**
true 1	169	1204

Table 3: Confusion matrix: a great number of false positive

Lastly, we focus on identifying the group of people are most likely to convert. Specifically, within the group that are correctly classified as unwilling to go regular check-up (labelled 0, $P(y = 1) \leq 0.5$), people with a higher predicted probability ($P(y = 1)$)) should be targeted since a slight change in covariates might flip their decisions. The histogram of the predicted probability is showed in Figure 6. It is observed that 40 people are within the range that $P(y = 1) \in (0.45, 0.5)$, while only a small proportion of people lie in the range of $(0, 0.15)$. some dominant patterns have been discovered by visually inspecting the data. Within the range of $(0.45, 0.5)$, most of the people have a higher degree, 80% of them have a health insurance, 85% of them believe health is important, 80% don't receive long-term treatment, and 80% think it is suitable to do exercise often (i.e., 40 minute per day).

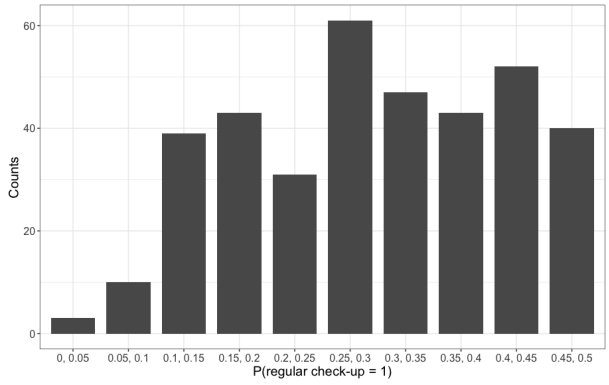

Figure 6: About 40 people are potentially easy to convert

Conclusion

To summarize, in this project we are investigating patterns and developing classification model that would be helpful for decision makers to refine the advertising strategies for regular check-ups in less developed countries. Using a survey dataset that contains 49 variables collected from 2068 respondents in Vietnam, we have discovered some interesting insights, and come to a set of suggestions to make such medical service more attractive to a broader range of people.

Firstly, we have found that most of the people have a neutral opinion on the skillfulness of the medical staff (*check-up quality*) as well as the general enticement of the information provided during check-up (*check-up information*). However, compared with the latter, people are more satisfied with *check-up quality*, especially on the doctor's d knowledge to assure professional reliance. On the contrary, there is still space to improve the attractiveness of the *check-up information*. In addition, since a high correlation have been observed within both of the feature set, it is highly skeptical that the respondent can truly differentiate these terms well. As a result, however, some of the ratings might not reflect the actual evaluation that the researchers plan to collect.

Furthermore, similar important factors have been identified for people with/without symptom of disease when making a decision for regular check-up. Specifically, using the impurity as importance

score in random forest, it is found that age, BMI, job status and amount of time suitable for exercise rank top 10 for both of the cases. Also, while *check-up quality* are more crucial for people without symptom, for people with potential illness the quality of *check-up-information* matters as well. Oppositely, less attention should be focused on factors including whether the respondent or family is in good health, takes simple medical measurements regularly, or has little faith in the quality of medical service.

Finally, a logistic regression model is implemented to not only predict whether a person would conduct check-up regularly, but also how likely a person can be persuaded to change his/her mind. The model can achieve a 0.82 prediction accuracy in terms of ten fold cross validation. Using the predicted probability as metrics, 40 out of the 369 people that are correctly classified are identified as potential customers to approach to. Although no unique pattern is extracted, having a higher degree, owning a health insurance, and working out often are positive signs for people that can be easily convinced for regular check-ups.

However, there are several limitations that need to be tackled in next stage. Firstly, since random forests are biased for variables with more levels, using the impurity score as importance measure might not be the best option since most of the variables in this dataset are categorical. On the other hand, another analysis should be conducted to reveal a common pattern of the people that can be easily persuade.