**Capstone project-1: Prudential Life Insurance Milestone Report**

**Contents**

1. Problem statement- why is it useful to answer the question
2. Clients and intended audience
3. Dataset used for the investigation
4. Data cleaning and wrangling
5. Data visualization
6. Exploratory data analysis (EDA)
7. Machine learning algorithms

**Problem statement- why is it useful to answer the question**

Insurance policies give a sense of security to the policy holder in case of uneventful things in life such as death, disability. At the same time the policy issuer have to make sure they do not end up losing money if there are more claims paid out  than the premiums collected for insuring people. This study/ model will use various data analysis techniques to predict the risk of insuring each customer based on risk factors such as age, weight, height, BMI, sex, medical history, employment history etc. The analysis also will address machine learning techniques to check the goodness of fit between the independent variables and the target variable.

**Clients/ Intended audience**

The study is of interest to any companies that offer insurance or loans to customers based on customer profiles to predict risk. The study already has classified the customers based on risk and this study works as a sanity check to improve model efficiency. Therefore, it can be applied to a variety of predictive analytics cases.

Even though this exercise involves term life insurance case, many other companies can benefit from this analysis, such as banks that offer other insurances, loans to consumers, health insurance market place and can be extended to other types such as home or auto insurances.

The analysis helps them to decide what types of customers are more risky to insure and come up with a model to predict insurance premiums based on risk.

**Dataset used for the investigation**

The dataset for this analysis comes from Prudential Life Insurance and attached is the link to the dataset.

<https://www.kaggle.com/c/prudential-life-insurance-assessment/overview>.

Since the data consists of private and sensitive information, many of the data labels (columns) are coded to make it ambiguous. For example, there are a list of variables containing applicants’ health information. But instead of listing the actual variable label such as ‘High blood pressure’, it may be listed as ‘Medical\_History\_1’. Similarly, a salary of the applicant can be listed as ‘Employment\_History\_1’. After the analysis we will only know if which coded variables affect the analysis. Prudential insurance company will take the analysis and match these coded variables to the actual variables affecting the results.

However, some variables such as age, BMI, Wt, Ht are provided without masking them. But these variables are normalized to a scale between 0 and 1.

The following table summarizes all the variables:

|  |  |
| --- | --- |
| Variable | Description |
| Id | A unique identifier associated with an application. |
| Product\_Info\_1-7 | A set of normalized variables relating to the product applied for |
| Ins\_Age | Normalized age of applicant |
| Ht | Normalized height of applicant |
| Wt | Normalized weight of applicant |
| BMI | Normalized BMI of applicant |
| Employment\_Info\_1-6 | A set of normalized variables relating to the employment history of the applicant. |
| InsuredInfo\_1-6 | A set of normalized variables providing information about the applicant. |
| Insurance\_History\_1-9 | A set of normalized variables relating to the insurance history of the applicant. |
| Family\_Hist\_1-5 | A set of normalized variables relating to the family history of the applicant. |
| Medical\_History\_1-41 | A set of normalized variables relating to the medical history of the applicant. |
| Medical\_Keyword\_1-48 | A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application. |
| Response | This is the target variable, an ordinal variable relating to the final decision associated with an application |

The response variable is an ordinal variable categorized 1 thru 8 and relates to the final decision associated with the application. Again, we will not know exactly what each of these categories mean.

**Data cleaning and wrangling**

* Data cleaning: The data is taken from Kaggle competitions and is provided by Prudential Life Insurance. The dataset is pretty clean and the variables are already arranged as columns. The variable names are coded to maintain privacy of the information. All the continuous variables are normalized.
* Missing values: There are variables with missing values. Some of these variables are continuous and some are discrete and categorical. Used the pandas data-frame method to fill the missing values in the columns:

**data = df.fillna(method='ffill').fillna(method='bfill')**

This method to replace all missing values with forward fill and backward fill data.

* Outliers: There is significant outliers present in some variables. Variables ‘Age’, ‘BMI’ and other continuous variables have some outliers. After plotting the response vs the variable of interest as bar chart, the effect of outliers seemed insignificant. Also, the outliers appear to be good data points rather than outliers. Tried choosing data within 15-85 percentiles to reduce the outliers but the effect one the data is minimal. Another drawback with eliminating outliers is loss of nearly 8000 observations of data. Therefore, even though, removing the outliers helps with data visualization, all the data is kept as is for the analysis.

**Data visualization**

|  |  |
| --- | --- |
| **Response** |  |
| **1** | 6207 |
| **2** | 6552 |
| **3** | 1013 |
| **4** | 1428 |
| **5** | 5432 |
| **6** | 11233 |
| **7** | 8027 |
| **8** | 19489 |

Following plots provide the visualization of the data for analysis. The continuous variables are plotted in histograms to check for normality. Other plots check the effect of some of the independent variables on the ‘Response’ variables grouped by the eight response categories.

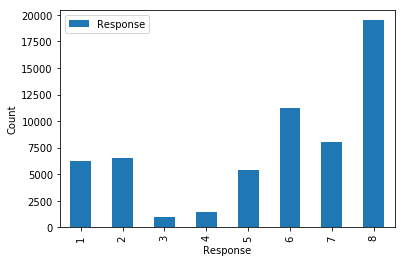
****

Fig-1: Response variable counts for the eight categories

The above figure demonstrates the response categories for the eight decisions taken on the applications.

|  |  |
| --- | --- |
|  | **BMI** |
| **count** | 59381.000000 |
| **mean** | 0.469462 |
| **std** | 0.122213 |
| **min** | 0.000000 |
| **25%** | 0.385517 |
| **50%** | 0.451349 |
| **75%** | 0.532858 |
| **max** | 1.000000 |

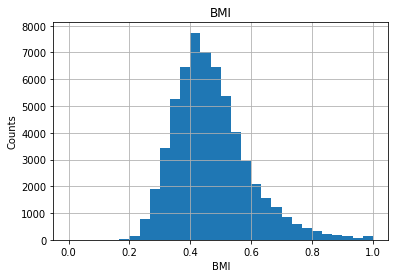
****

Fig-2: BMI distribution and summary statistics

The BMI histogram is shown here as an example to continuous variables with normal distribution. Other variables such as age, height, weight, and some employment info variables (most likely salary info) show normal distribution.

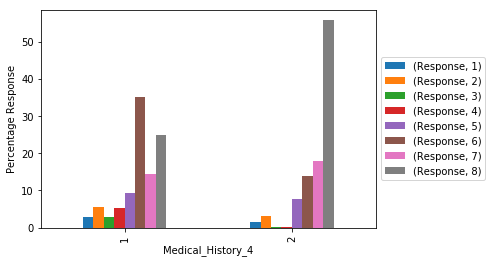
****

Fig-3: Categorical variable examples Medical\_History percentages of Response

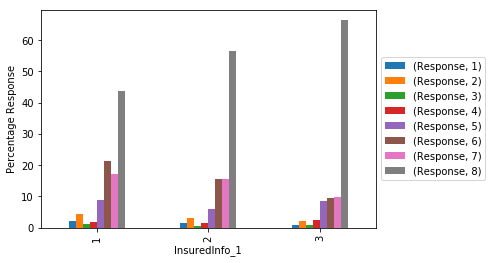
****

Fig-3: Categorical variable examples InsuredInfo\_1 as percentages of Response

The above two figures represent the data distribution for categorical variables. Most of the variables are categorical. Again, as explained earlier, with labeling ambiguity it is hard to classify each categories. However, the effect on the eight categories as a percentage is shown in the above two plots.

**Exploratory Data Analysis (EDA)**

**Are there variables that are particularly significant in terms of explaining the answer to your project question?**

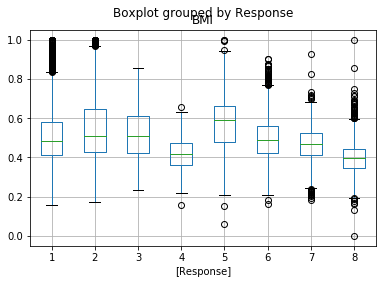
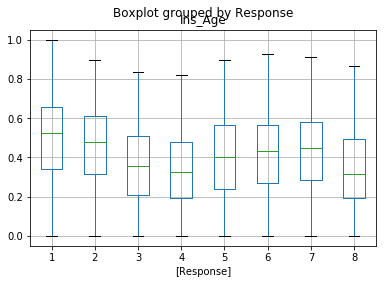
The ‘Response’ variable is the final decision associated with the insurance application. From the problem definition we can think of certain variables that can affect an insurance application. Some of the most significant ones are: BMI, age of the applicant, medical history of the applicant, employment status and family history of the applicant.

**Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?**

Yes some of the independent variables have strong correlation between them. For example, BMI of an applicant is calculated with a person’s height and weight and we see a strong correlation between them using correlation coefficients between them. Checked also the correlation between the ‘Response’ variable and some of the independent variables.

**What are the most appropriate tests to use to analyze these relationships?**

* Visual analysis using the box plots to see the response variable vs the independent variables give us information about the relationships. Below we have two plots with independent variables Age and BMI of the applicants grouped by dependent variable ‘Response’.



* The correlation matrix to check dependency between variables is done to see the effect of some of the variables on the dependent variable. Ran a correlation matrix to check some of the independent variables effect on the response variable. As summarized in the table below, the person’s age, BMI, weight are all inversely correlated to the response. As expected, with increase in age, weight, BMI of applicant, the approval goes down.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Ins\_Age** | **BMI** | **Ht** | **Wt** | **Response** |
| **Ins\_Age** | 1.000000 | 0.137076 | 0.008419 | 0.110366 | -0.209610 |
| **BMI** | 0.137076 | 1.000000 | 0.123125 | 0.854083 | -0.381601 |
| **Ht** | 0.008419 | 0.123125 | 1.000000 | 0.610425 | -0.093576 |
| **Wt** | 0.110366 | 0.854083 | 0.610425 | 1.000000 | -0.351395 |
| **Response** | -0.209610 | -0.381601 | -0.093576 | -0.351395 | 1.000000 |

* Significance tests to check for p-value to see if the data is significant: The dataset has a total of 59300 data points. The sample size is large enough to be statistically significant. Also, conducted analysis of variance for f-stats and probability value. The null hypothesis (H0) is that the data is not statistically significant. Analysis of some of the important variables indicate the p-value = 0 which is less than alpha=0.05 and therefore reject the null hypothesis.

Also, did multiple comparisons of means between the response categories vs the independent variables such as applicant age, BMI etc. Since P-value = 0 we can reject the null hypothesis that difference in means between categories of Response = 0.

The data below shows the comparison of means for ‘BMI’. As we can see, the mean difference is significant at alpha = 0.05 and therefore we can reject null hypothesis that there is so difference between the means of the response categories. Therefore, the data is significant.

