**Capstone project-1: Prudential Life Insurance Exploratory Data Analysis (EDA)**

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

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 |

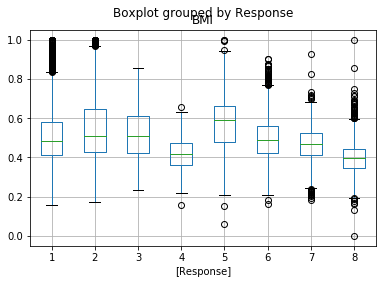
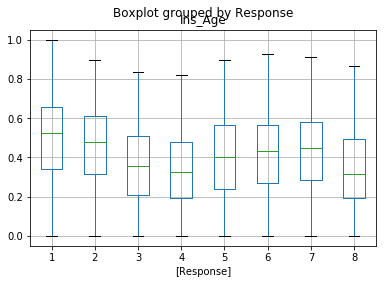
The ‘Response’ variable is the final decision associated with the insurance application. From the problem definition we can think of certain variables that can affect an insurance application. Some of the most significant ones are: Age of the applicant, medical history of the applicant, employment status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

1. 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’.



1. 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 |

1. 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.