Basic Data Analysis for Kaggle Prudential dataset

The Kaggle Prudential set consists of 59,381 observations and 127 categories. The goal of the competition is to use the dataset to build the 'best-fit' model that predicts the response category using the 127 characteristics.

The descriptive variables can be put in three rough classes: Demographic, product and insurance, and medical history. Product and insurance includes information revealed by the applicant, in terms of which product they have applied for, and information about past assessments and choices by the applicant. The demographic information contains both personal (age, weight, height, bmi), employment info and family history. These are clearly independent variables. The third class is medical history, this includes both the medical history and medical keywords categories.

Below I will pull out some of the results, and talk about some of the potential data issues.

In the data, the response variable, or risk category, varies from 1 to 8. Below I've shown the distribution of the observations across risk categories:

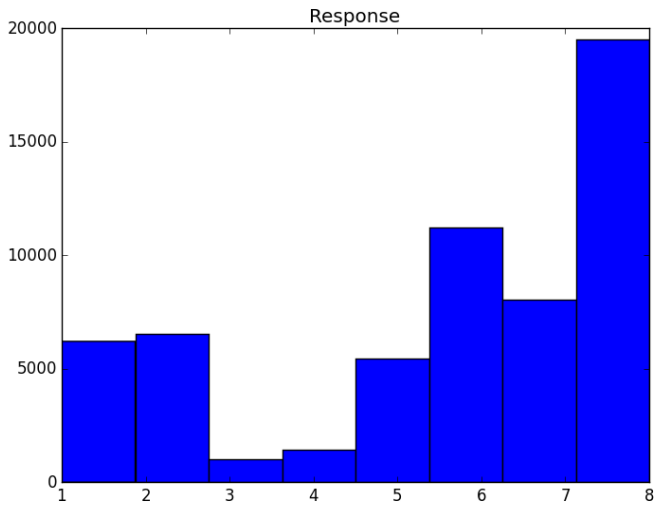
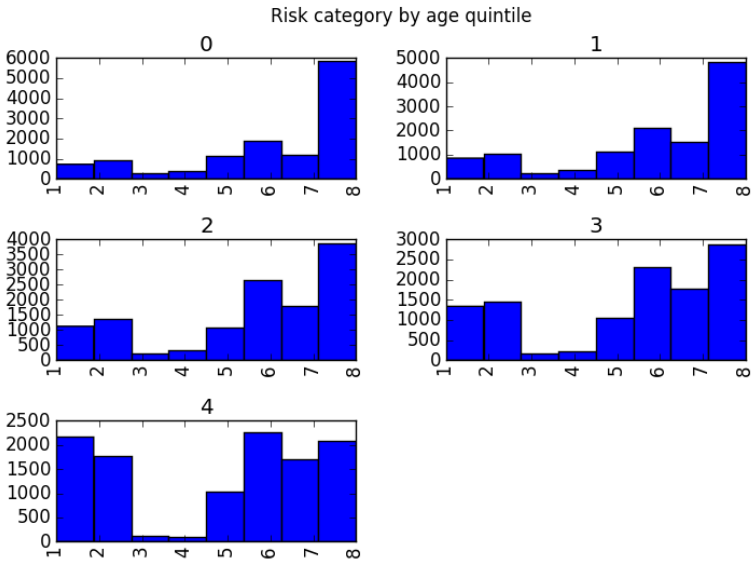


Figure 1: Histogram of response variable

A couple of things to note about the distribution across risk categories. First, it is very bi-modal. It seems that the risk levels are either high, or low, with very few in the middle. Unless the underlying population is bi-modal, this suggests that the categories might not be linear, and a linear predictive model may be a bad fit for the data. There are very few observations in risk categories 3 and 4, so the best fit with a linear predictive model would likely be to make the intervals for those categories quite small.

Demographic Variables

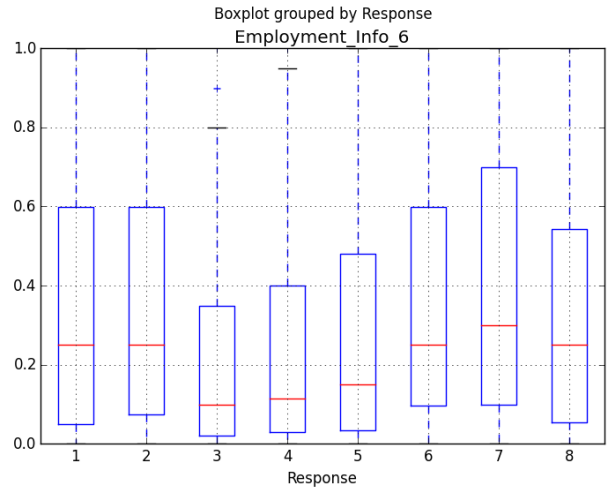


When we break down risk categories by age quintiles we find a slightly different pattern.

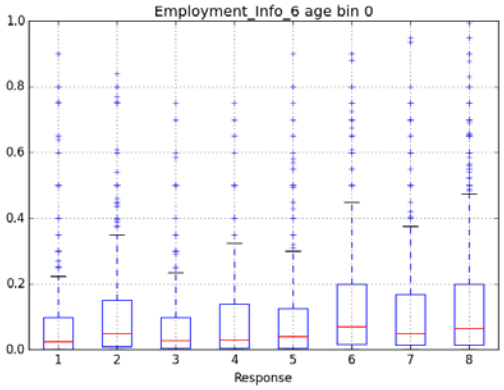
Figure 2: Histogram of Response Variable by Age

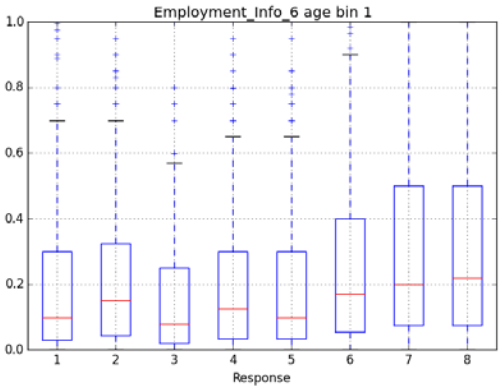
Assuming that the age variable goes from youngest (0) to oldest (1), the data suggests that the youngest quintile are most likely to be in risk category 8, while the oldest quintile seem split between low and high risk categories, with large numbers at each end. The frequency with which an observation is put in categories 1 and 2 seems to be increasing with age. It seems likely that the price of insurance is calculated based on both age and risk metric, as the metric alone doesn't seem to take into account the full impact of age.

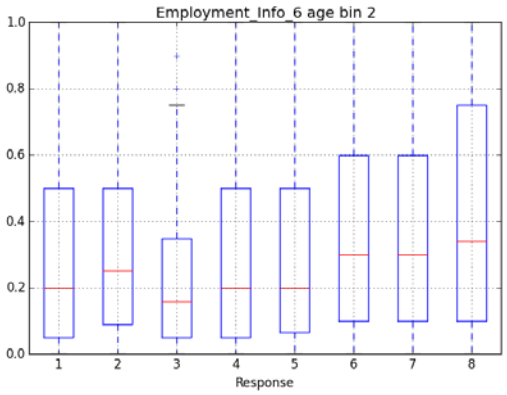
When looking at employment info variables it is important to take age into account. For example, looking at Employment_Info_6, the boxplot across risk categories looks as follows:

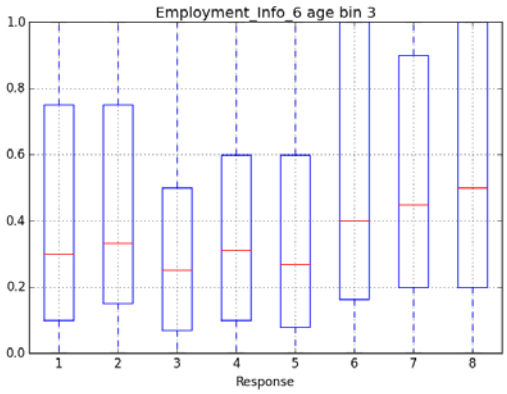


When we separate it out into the five age quintiles, we get the following graphs:









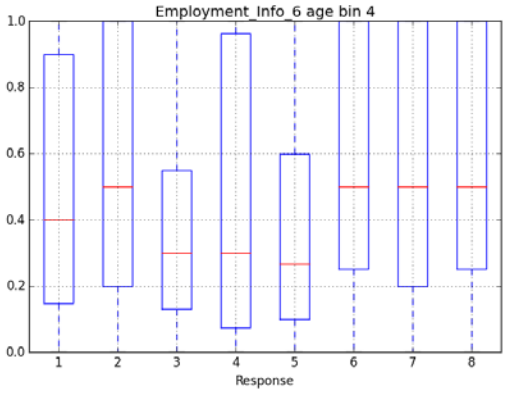


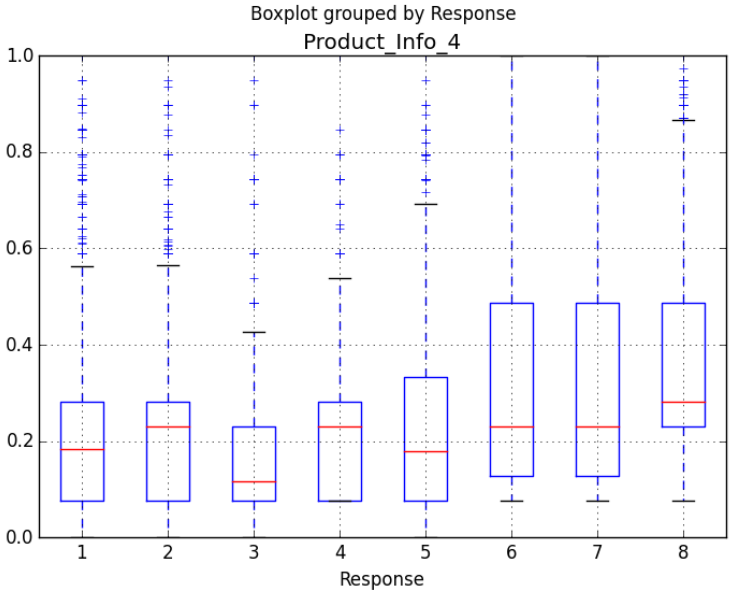
Figure 3: Employment info 6 by age quintile

The difference between the lowest and highest age quintiles is stark. Simply using employment info 6 without accounting for age would result in employment info 6 acting as a proxy for age, and muddying any analysis on the effect of the variable itself. It also appears that the data is truncated at 1.

The intuition behind why we want to take age explicitly into account when looking at these variables is as follows. A young person might have a limited employment history simply because they are young, whereas for an older person a limited employment history may be a signal of underlying health problems, or some high risk factor. Even if we include age as a variable, not taking into account the interaction between age and employment info obscures the real interaction between employment history and risk.

Product and Insurance

The only continuous variable for product information is Product_Info_4. We can see below that the risk category seems to be increasing with the variable:



Appendix 1: Python Code:

```
from numpy import loadtxt, zeros, ones, array, linspace, logspace
from pylab import scatter, show, title, xlabel, ylabel, plot, contour
import csv
import pandas as pd
import matplotlib.pyplot as plt
traindf = pd.read_csv("train.csv")
traindf.head(n=10)

#Histogram for response variables to get an idea of distribution of response
traindf.hist(column='Response', grid=False, bins=8)

#Create age bin column for age quintiles
traindf['Age_Bins']=pd.qcut(traindf['Ins_Age'], 5, labels=False)

#Plot the data for Response variables by age
axes = traindf['Response'].hist(by=traindf['Age_Bins'], bins=8, grid=False)
plt.suptitle('Risk category by age quintile')
plt.show()

#Plot the data for float variables across all ages

for i in range(1,128):
y=traindf.iloc[:,i]
yname = y.name
if y.dtype == 'float64':
traindf.boxplot(column=yname,by='Response')
show()

#Plot the data for float variables by age

for j in range(0,5):
traindf_age = traindf[traindf.Age_Bins == j]
for i in range(1,128):
y=traindf_age.iloc[:,i]
yname=y.name + ' age bin ' + str(j)
if y.dtype == 'float64':
traindf_age.boxplot(column=y.name,by='Response')
plt.title(yname)
plt.suptitle('')
show()
```