Data Manipulation and Feature Engineering – Prudential Insurance Kaggle dataset

The variables in the dataset can be put in three categories:

Demographic:

Description Height, Age, weight, BMI, Family History

Variables:

Continuous: Ht, Ins_Age, Wt, BMI, , Family_Hist_2, Family_Hist_3, Family_Hist_4, Family_Hist_5

Categorical: , Family_Hist_1

Product and Insurance:

Description: Product applied for (Product_Info), Insured information

Continuous: Product_Info_4

Categorical: Product_Info_1, Product_Info_2, Product_Info_3, Product_Info_5, Product_Info_6, Product_Info_7InsuredInfo_1, InsuredInfo_2, InsuredInfo_3, InsuredInfo_4, InsuredInfo_5, InsuredInfo_6, InsuredInfo_7,

Historical variables:

Continuous: Insurance_History_5, Employment_Info_1, Employment_Info_4, Employment_Info_6

Categorical: Medical_History_2, Medical_History_3, Medical_History_4, Medical_History_5, Medical_History_6, Medical_History_7, Medical_History_8, Medical_History_9, Medical_History_11, Medical_History_12, Medical_History_13, Medical_History_14, Medical_History_16, Medical_History_17, Medical_History_18, Medical_History_19, Medical_History_20, Medical_History_21, Medical_History_22, Medical_History_23, Medical_History_25, Medical_History_26, Medical_History_27, Medical_History_28, Medical_History_29, Medical_History_30, Medical_History_31, Medical_History_33, Medical_History_34, Medical_History_35, Medical_History_36, Medical_History_37, Medical_History_38, Medical_History_39, Medical_History_40, Medical_History_41, Insurance_History_1, Insurance_History_2, Insurance_History_3, Insurance_History_4, Insurance_History_7, Insurance_History_8, Insurance_History_9, Discrete: Medical_History_1, Medical_History_10, Medical_History_15, Medical_History_24, Employment_Info_2, Employment_Info_3, Employment_Info_5, Employment_Info_1, Employment_Info_5, Medical_History_32

Medical_Keyword_1-48 are dummy variables.

Feature Normalization:

Continuous variables were checked and modified if necessary to ensure they each had a mean of zero and a standard deviation of 1.

If a continuous variable was truncated, such that the 75th percentile is equal to the max, it may be changed to a discrete variable (i.e. less than x and more than x) in which case its name was changed to oldname_x, which has a value of 1 if the variable is greater than x, and 0 otherwise.

Feature Engineering:

For each categorical variable, new dummies are created such that each category i with more than 10 observations is given a new name, oldname_i which is 1 if the category for that observation is i, 0 otherwise. Any remaining observations are labelled as oldname_x (i.e. any categories that have fewer than 10 observations). This was done in order to keep outliers from skewing the coefficient for any individual category.