Capstone Report

Monisha Gopal

11/21/2017

Introduction

BNP Paribas Cardif is an global insurance company that specializes in personal insurance. Not only do they have to deal with an increasing number of claims, their customers are also expecting them to handle claims as fast as possible. Usually claims go through a number of checks before being approved. However, BNP Paribas Cardif hopes to speed up that process using data science.

Data

BNP Paribas Cardif provided an anonymized dataset on kaggle.com with two categories of claims described on kaggle.com as follows: 1. claims for which approval could be accelerated leading to faster payments 2. claims for which additional information is required before approval

The goal is to predict the category of a claim based on information available early in the claims process.

Initial Look at Data

Three files are provided: \begin{enumerate}

Part I

t

rain.csv - training set with target (dependent variable)

Part II

t

est.csv - test set without target

Part III

s

ample_submission.csv - sample submission with correct format \end{enumerate}

The dataset contains 133 features named ‘ID’, ‘target’, and ‘v1’ through ‘v131’. There are both categorical and numerical features. None of the categorical features are ordinal (specified on kaggle).

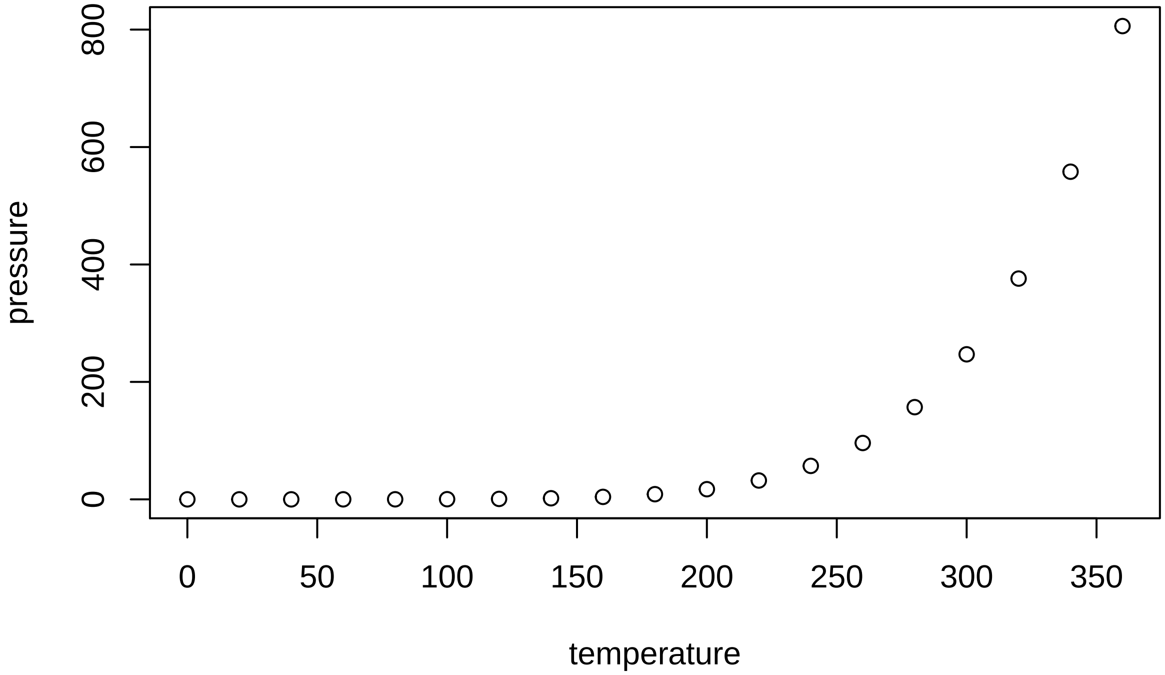
The main limitation of this dataset is that the features are anonymized. Because of this, we can’t use domain knowledge to eliminate features or predict the distributions of certain features.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.