Exploratory Analysis of Dataset

Description

The given dataset contains 300 observations and 44 columns in which last 30 numerical columns seems to be sensor feeds from 30 sources. The 1st numerical column 'Y' seems to be the predicted variable. Column 'ID' is unique a ID and didn't taken for any visualizations. The column 'Date' consists of date range in 'yyyy-mm-dd' format. Remaining 11 columns contains different nominal and ordinal factors. Columns 'Price', 'Priority', 'Speed', 'Duration', 'Temp' seems to be ordinal while columns 'Surface', 'Class', 'State', 'Agreed', 'Location','Author' seems to be ordinal. The ordinal columns will be converted to factors with levels to get deserved orderings. 'Date' column is converted to normal date type using 'lubridate' library so could be used for time series and similar visualizations. Extraction of months from Date column may give a new cyclic variable.

Procedure

1. **Summary**

In this analysis, the summary statistics and characteristics of data is observed with the help of summary and glimpse functions in R. From summary, apart from columns ' Y, ID, Author, Date, Priority' NAs are observed and column 'sensor7' has highest number of NAs(66). The NA observations aren't resolved because of the lack of domain knowledge. From the mean and median values skewness is identified. Variables 'sensor3', 'sensor4', 'sensor13', '17', 'sensor22', 'sensor24', 'sensor27' have very high difference in mean, median values.

2. **Visualization – Missing Values**

Missing values in the dataset was plotted with the help of 'vis_miss()' method. Apart from column 'sensor7' missing values in other columns seem to be tolerable.

3. **Visualization – Categorical**

a) Mosaic Plot

The mosaic plot represents the frequencies of each factor in a categorical variable against another categorical variable. The area of mapped region corresponds to the occurrence of factors together which is distinguished by p-values. Comparing 'Duration' and 'Author' factor "Very Long" in 'Duration' and "XX" in "Author" are least found together. Similarly, factors "Medium" in 'Priority', 'No' in 'Agreed' and 'Checked' in 'State' also has very low p-value. The plot also supports multiple column inputs and is responsive.

4. **Visualization – Numerical**

a) Box Plots

Box plots are used to identify outliers, skewness, range of quantitative variables. With standardization at 1.5*IQR, it can be seemed that columns 'sensor3', 'sensor4', 'sensor13', 'sensor17', 'sensor22', 'sensor24', 'sensor27' have extreme high outliers which needs explanation from the expert confirms findings from summary tables. As well as these columns are very less varied throughout the observations while other columns are more spread. Apart from that from the box plot ‘sensor7’ seemed to be left-skewed. The variable ‘Y’ has really large whiskers means that minimum and maximum values are really far from Q1 and Q3 respectively. For getting more outliers an IQR multiply(<1) is preferred.

b) Correlogram Plot

Correlograms help us visualize the data in correlation matrices. The color coding from Blue to Red implies correlation from high to low(+ve to -ve). Correlation between ‘Y’ and sensors are like, ‘sensor3’, ‘sensor4’, ‘sensor13’, ‘sensor17’, ‘sensor24’, ‘sensor27’, ‘sensor22’, ‘sensor23’, ‘sensor21’, ‘sensor25’, ‘sensor28’, ‘sensor29’, ‘sensor29’, ‘sensor26’, ‘sensor30’ are negatively correlated with ‘Y’ while others are positively correlated with ‘Y’

c) Pair Plots

Pair plots also show the correlation between variables in a more detailed way including correlation values. Relationship plots against each variable are obtained as detailed scatter graphs here. Majority of sensor variables has no visible linear relationship with expected predicted variable ‘Y’ while some of sensor variables have a linear relationship with each other.

d) Rising Order Chart

Rising Order Chart is used to find the discontinuity in dataset. With scaling we could see that columns 'sensor3', 'sensor4', 'sensor13', 'sensor17', 'sensor22', 'sensor24', 'sensor27's have a sudden discontinuity after 270 observations rapidly on upward direction. This may be due to change measuring units or change in environment.

e) PCA Chart

Principal component analysis is used for multivariate problems to reduce dimensionality and makes it linearly comparable. PCA of numerical variables(sensor1-30) can’t be used to find novelties, because there is no proper clustering or distinction.

f) Time Series Chart

Time series chart provides visual representation of numerical variables over the given time period and is used analysing parameters like trend, seasonality etc. Here for numerical columns ‘sensor3', 'sensor4', 'sensor13', 'sensor17', 'sensor22', 'sensor24', 'sensor27' values got sudden upward shift shortly after year 2006.