5 Discussion Section

In this section we discuss any observations you noted and any recommendations you can make based on the results.

5.1 Selection of machine learning models

We used the following machine learning models for car accident prediction:

- K-Nearest Neighbors (KNN)
- Decision Tree

5.2 Accuracy of the models

The obtained accuracy of both seems to be suitable for resolving the problem of car accident prediction on current available weather observations and conditions for a planned car travel.

5.3 Technical Installation Prerequisites

For the models we need a number of python libraries. To install the right version of these in the development environment without having any python error messages when developing the models was a tedious job.

For KNN the necessary libraries could be installed and a graphical plot could be printed.

For Decision Tree the necessary libraries most could be installed. We experience timeouts when installing the libraries for Decision Tree and a graphical plot could not be printed. This problem needs to be fixed in the near future.

5.4 Unresolved Error Messages

The following error message when preparing the decision tree plot could not be resolved:

InvocationException: GraphViz's executables not found

5.5 Integrated Development Environments IDEs

We used to different IDEs:

- IBM Cloud, Jupyter Notebook

- Anaconda on iMac OSX Jupyter Notebook

Both work fine for the development of the machine learning models

Setup the specific technical environment for Jupyter could be improved for machine learning problems.

Elapsed installation times for prerequisite python libraries in both IDEs are not acceptable long.

5.5 Other machine learning models

5.6 Business Understanding phase
The initial phase to understand the project's objective from the business or application perspective could be resolved.

Translation of this knowledge into a machine learning problem with a preliminary plan to achieve the objectives could be resolved.

5.7 Data understanding phase
Collecting or extracting the dataset from various sources such as csv file or SQL database could be resolved. csv File was used for that task. A SQL database dis not apply.

Determining the attributes (columns) that are useed to train the selected machine learning model could be resolved. Also, assessing the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on was resolved initially, but could be improved. At leaset we found only two severity categories, which depend not only on the current weather conditions. These patterns need further studies.

5.8 Data Preparation phase
Data preparation included all the required activities to construct the final dataset which were fed into the selected modeling tools. Data preparation was performed multiple times and it included balancing the labeled data, transformation, filling missing data, and cleaning the dataset. This was until now the major effort of the project: data cleansing. The available data quality needs in general more attention, which can not be assumed without effort. This is also true for the right data available for statistics problems to be resolved with the selected models.

5.9 Modeling phase
In this phase, in general various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. We selected only two: KNN and decision tree. Furthermore, SVM, XGBoost, decision tree, or any other techniques could be selected as well. This is for further studies. In general, a single or multiple machine learning models for the

same data mining problem could be selected. At least, herewith only two machine learning models were selected. At this phase, stepping back to the data preparation phase was often required. This was also high effort prone.

5.10 Evaluation phase

Before proceeding to the deployment stage, the model needed to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics could be used for the model evaluation such as accuracy, recall, F1-score, precision, and others. For this project, only accuracy was calculated ans evaluated for the selected machine learning models. Both have acceptable values. Other metrics are for further studies.

5.11 Deployment

In general, as he deployment phase requirements varies from project to project, the report is deployed to a website of the author. As this can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment, the working files a submitted to the authors' Git hub. In this environment, the possible customers or end-users can utilize the model in different ways such as API, website, or so on. At least, this work is published to everyone interested.

Published as a blog on: www.energing.de

Guthub: https://github.com/herbfrisch/jbs_cornerstone