Pilot Study for
CE802 – Machine Learning & Data Mining

Assignment

Design and Application of a Machine Learning System for a Practical Problem

Professor - Dr Luca Citi

Student Name - Karamjot Singh

Registration number(s): 2006510

Date - 20 January 2021

Words count: 660

Problem Statement

The goal is to make a system for a travel insurance company to predict future claims

Proposal

Nowadays, companies have a dilemma of selecting between traditional programming and a machine learning approach towards the solutions for various problems. To resolve this problem, they first have to understand each approach's benefits and limitations before implementing them.

The following table gives a general comparison between traditional programming and machine learning.

Approach	**Rule-based**	**ML-based**
Sub-Goal	To recognize parent-child relations based on event attributes.	
Operation	Rule generation, rule-based identification	Feature extraction, train & test
Model	Empirically derived rules	Automatic generated
Labeled Data	No	Yes (for training)
Manual Effort	Needed for generating rules	Minimal
Major Cost	Rule tuning by human	Pairing operation
Flexibility	Low (cannot recognizes beyond rules)	High (can adapt to subtle cases)

In our case, customer filing a claim on their travel could rely on many factors or patterns that are unknown or can't be seen easily. There is no way to write a rule-base by a human for such problems which are influenced by several independent features, with many of these rules potentially overlapping or frequently requiring tuning. On the other hand, supervised machine learning can handle several features by mapping them to the target variables based on the example dataset. That's why machine-learning procedures are more feasible as compared to rule-based programming for our problem.

Problem Type

Predicting a class of Target variable is a classification problem as the outcomes fall under one of the predetermined categories i.e. True or False (True when insured files claim and False when not).

The historical data of past policies with the information of the insured is given by the company. Since it is a classification problem where the labelled dataset is provided with the target class, the learning algorithm to be reviewed for the research on this data are:

- Decision Tree
- Support Vector Machine
- Logistic Regression
- Neural Network
- Ensemble Learning Techniques
    - Boosting
    - Stacking
    - Bagging

There are factors related to data to select a perfect algorithm for our task, for example:

- **Dimensionality of Data**: the speed of execution of some algorithms affected by the size of data
- **Features of Data**: some algorithms are very sensitive to the scale of the numerical values
- **Type of Data**: if the target variable is categorical, classifiers are used and if numerical, regressors are used

Performance Evaluation

The best way to evaluate a machine learning model is to divide the dataset into training and validation set and starts working on different models with the basic approach. After getting feedback from various performance metrics such as the confusion matrix, classification report, learning curves, validation curve, accuracy on the validation set, the model would be trained until the desired results achieved. For better visualization, plots would be made like histogram, pairplot and barplot for a deeper understanding of the data.[2]

Feature Suggestion

Travel insurance has many different components, for example, covers for illness, accidental injuries, trip cancellation, lost or damaged possessions, emergency evacuation and many more. The list is endless and therefore also the number of features on which the prediction relies.
As asked, the following are some features that may be good predictors used for classification:

- **The Age of the Insured**: the more the age, the more the insured is susceptible to illness
- **Pre-Existing Medical Conditions**: insured travellers suffering from a medical condition are more vulnerable to file a claim in the future
- **Frequency of Travel**: more the travel more the chances of filing a claim
- **Adventurous activities while travelling**: dangerous activities like skiing and sky-diving may increase the risks of accidents
- Whether a claim filed or not
- **Duration of travel**: the longer the insured travels the more likely the risk associated with it, for example, accidents and thefts
- **What's going on in the world**: like terrorism or pandemic

To summarize, the price of any insurance premium depends upon the risk factor related to the insured situation.

Conclusion

This study concludes the advantages of machine learning approach over rule-based programming for the given problem. The above-mentioned learning procedures will be considered for predictions, however, the 'no free lunch' theorem in machine learning states that "**no single machine learning algorithm is universally the best-performing algorithm for all problems**", it is difficult to finalize an optimal model before analyzing and evaluating the performance of each algorithm on the data.[1]

References

1. *What "no free lunch" really means in machine learning* (2020). Available at: https://towardsdatascience.com/what-no-free-lunch-really-means-in-machine-learning-85493215625d (Accessed: 11 January 2021).
2. *Evaluation Metrics Machine Learning*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ (Accessed: 15 January 2021).
3. *Table 1: Comparison of properties between rule-based and machine...* (2021). Available at: https://www.researchgate.net/figure/Comparison-of-properties-between-rule-based-and-machine-learning-ML-based-approaches_tbl2_291419651 (Accessed: 16 January 2021).