PILOT STUDY

(Word Count:770)

(Reg Id: 2003258)

Objective

A travel insurance company would like to offer a discounted premium to customers that are less likely to make a future claim. The aim of the project is to able to predict whether an insured person makes a claim or not with high degree of certainty using past polices and insured information. Discuss the nature of the problems and models that can be used for the problem and evaluation methods to identity most suitable model before deploying.

Task

The data available for a model are user information, policy data, purpose and destination of travel and whether an insured claimed an insurance or not. So, it is clear we have a multiple input features and output or target variable, that is whether was insurance claimed or not, so supervised learning algorithms can be used for this problem.
Here, target variable can take two classes i) insurance claimed ii) insurance not claimed, hence it is a binary classification problem

Features

Some of the features that would be helpful to make right prediction are as follows:
Age, Sex, Destination, purpose of visit, Duration of stay, travel agency, travel agency type, policy product, HasAnyMedicalHistory, credit ratings in financial system (if possible),
Name (or a unique Id which would be helpful in pre-processing to get information like if a person has made any claims before, if made, how many?)

Learning Procedure

Logistic regression: It works well with simple data sets and performs well if data is linearly separable. It is very easier and efficient to train. Although feature scaling make convergence faster it is not mandatory to do it.
It is easy to interpret, just by observing coefficient of features, we can understand how important corresponding features are and the direction they are taking the decision.

SVM: Unlike some of the learning algorithms, SVM does not require data to be linearly separable, its can solve complex relations.
it tries to find the best margin (distance between the line and the support vectors) that separates the classes and thus reduces the risk of error on the data and make it not prone to outliers. And, SVM reduces risk of overfitting.

Decision Tree: Compared to other algorithms, DT's do require much pre-processing, data need not to be scaled. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders. Compared to other variants like Random forests it is computationally cheap. It is nonparametric method.

KNN: It is very easy and efficient to implement, particularly when number of features and training examples are not very high in number. The non-parametric nature of KNN gives it an edge in certain settings where the data may be highly unusual.
It has only 2 parameters to tune (K and distance function) and a good value of K makes it robust to outliers. KNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period. It stores the training dataset and learns from it only at the time of making real time predictions

Gradient Boosting: Although it is computationally expensive and needs longer to train, most of the time it gives better result than other models.

A model is a simplified representation of reality, and the simplifications are made to discard unnecessary detail and allow us to focus on the aspect of reality that we want to understand. These simplifications are grounded on assumptions; these assumptions may hold in some situations but may not hold in other situations. This implies that a model that explains a certain situation well may fail in another situation.

The “No Free Lunch” theorem states that there is no one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem, so it is common in machine learning to try multiple models and find one that works best for a particular problem. Depending on the problem, it is important to assess the trade-offs between speed, accuracy, and complexity of different models and algorithms and find a model that works best for that particular problem.

Evaluation metrics

The problem we are solving is a binary classification, so there are multiple evaluation metrics we can work with, some of them are:
i)Classification Accuracy
ii)confusion matrix
iii)AOC
iv)F1 score
Classification Accuracy can be used if target variable is evenly distributed, otherwise it may be not give true picture of goodness of model. In this particular problem the data is unevenly distributed , since we don't have to minimize FalsePositives or FalseNegatives,F1 score can be good metric. it gives an overall performance of model.

Reference

https://machinelearning-blog.com/2018/04/23/logistic-regression-101/

https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16#:~:text=Difference%20between%20SVM%20and%20Logistic%20Regression&text=SVM%20works%20well%20with%20unstructured,is%20based%20on%20statistical%20approaches.
https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a#:~:text=Compared%20to%20other%20algorithms%20decision,data%20preparation%20during%20pre%2Dprocessing.&text=A%20decision%20tree%20does%20not%20require%20scaling%20of%20data%20as,tree%20to%20any%20considerable%20extent.

http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html

https://chemicalstatistician.wordpress.com/2014/01/24/machine-learning-lesson-of-the-day-the-no-free-lunch-theorem/#:~:text=The%20%E2%80%9CNo%20Free%20Lunch%E2%80%9D%20theorem%20states%20that%20there%20is%20no,best%20for%20a%20particular%20problem.