Executive Summary:

Classifying "Y" in Insurance Data

U. Richard Iloanugo

May 25, 2022

1 Introduction

This report provides information on the predictive analysis for the classification coding exercise. The report discusses the data cleaning, the machine learning methods used, the results section and the conclusion.

2 Methodology

2.1 Data

The raw training data contains 40,000 rows and 101 features (including the target feature). Some of the features had a considerable amount of missing values; 42 features contained missing values with some missing as high as 32000 entries. The data engineering technique I used to resolve this issue was to drop features that had more than 70% of entries missing (3 features fell into this category). Afterwards, This reduced the number of missing entries in the training set, however, there was still a substantial amount of information loss. After excluding the 3 problematic features from the dataset, only 18 rows were left after I dropped rows with missing values. I applied an alternative method for dealing with large numbers of missing values, that is to replace missing values with the mean and mode of the features for continuous and categorical values respectively.

The next challenge with the data was how to deal with the non-numeric categorical features in the dataset. The categorical features in the dataset varied from dummy to multi-level categories. There are three main methods of dealing with this issue: one-hot encoding, frequency encoding and probability encoding. Feature encoding and probability encoding can result in data leakage in this case because I will use the entire distribution of the target and independent features to conduct the encoding (In the exercise, the training dataset will be used to train and validate the model before applying it to the test set). I employed the one-hot encoding technique even though this can generate a large number of additional features, but this was not an issue for this project. After the data engineering procedure, the training data had 177 features and retained all 40000 rows. The machine learning estimation strategy will split this data between training and validation data through cross-validation.

2.2 Machine Learning Models

2.2.1 Logistic Regression

Logistic regression is one of many classification models for machine learning. The logistic model uses a process of modelling the probability of a discrete outcome given an input variable. In a logistic regression classification problem, we are trying to determine if a new sample fits best into a category. This class of model is easy to implement and achieves very good performance with linearly separable classes. The logistic regression is essentially a transformation of linear regression using a sigmoid function, that is, it takes a linear combination of features and applies them to a non-linear sigmoid function. The model uses this method to predict the probability of a sample belonging to a class of a binary outcome. However, logit models can be extended to targets with multiple classes as with multinomial logistic models. I applied this method to the training data to set the parameters for predicting the probabilities of Y=1 in the test set.

Advantages of Logistic Regression

i Logistic regression is easy to implement and performs well in out-of-sample data compared to other machine learning models when the dataset is linearly separable.

Disadvantages of Logistic Regression

i The core assumption of logistic regression is that the data is linearly (or curvy linearly) separable in space. However, linearly separable data is rarely found in the real world.

ii The log odds output of logistic regression makes it difficult to quickly decide on appropriate targets.

2.2.2 Decision Tree Classifier

A Decision tree is a classifier that builds a flowchart like a tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The tree is created by splitting the training set into subsets based on the value test of attributes. The same process is repeated on each subset of the training set recursively. The recursion is completed when the subset at a node all has the same value of the target variable (pure leaf), or when splitting no longer adds value to the predictions. Therefore, an instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute. as shown. The process is then repeated for the subtree rooted at the new node. I use this machine learning method as an alternative to the logistic regression in this study.

Advantages of Decision Tree Classifier.

i Decision trees do not require data to be linearly separable, unlike logistic regression.

ii Decision trees can generate easily understandable rules for decision making.

Disadvantages of Decision Tree Classifier

i The decision tree can be computationally expensive to train.

ii Decision trees are prone to overfitting.

2.3 Analysis

The logistic regression analysis involved using the LogisticRegression() package from Scikit Learn on python version 3.9.7. The dataset was standardized using the mean and standard deviation of each independent feature in the model. I also used the grid search method to select the best hyperparameters for the logistic regression model. Note - These procedures were implemented after the data was split for training and testing to prevent issues regarding data leakage. The model performance was evaluated using the ROC-AUC metric.

The decision tree algorithm was implemented using the DecisionTreeClassifier() package from Scikit Learn. Other parts of the process were conducted similarly to the logistic regression model above.

3 Result Discussion

Table 1: **Classification Model ROC-AUC Score**

	ROC-AUC Score
Logistic Regression	0.7679
Decision Tree	0.7403

Notes: The table is the result of the ROC-AUC score for the classification model. We see from the table that both models performed almost in the same way. However, the logistic regression performs slightly better than the decision tree model.

The table above shows the performance score of the logistic regression and decision tree machine learning models. We can see that both machine learning algorithm has a similar prediction performance on the validation set. The logistic regression shows a performance of 76.79%, while the decision tree performed slightly less effective at 74.03%. Based on the ROC-AUC metric, the logistic regression model is better at predicting Y. Therefore, for this reason, and the fact that the logistic regression is easier to implement I advise our client is to use the logistic regression model in predicting new samples for Y.

4 Conclusion

The report provides information on the predictive analysis for the classification coding exercise. The training set contains 40000 rows and 101 features, including the main target feature. The method applied two classification algorithms: logistic regression classifier and decision tree classifier. After building and validating the learning models, logistic regression had a slightly better performance using the ROC-AUC scoring metric. Lastly, I use both machine learning models to make predictions with new samples from a training set. I conclude that our business partner should use logistic regression in predicting Y based on its superior classification performance in this exercise.