TECHKRITI IIT KANPUR ML/AI SUMMER INTERNSHIP PROJECT SYNOPSIS

Myself Harshvardhan Pardeshi, B. Tech Information Technology 3rd Year 6th Semester student from Guru Ghasidas University, Bilaspur (C.G.).

This file is the documentation for my project titled **“Insurance Claim Status Predictor”**. I created this project as it was assigned to me during my internship period at ***TECHKRITI’21 IIT KANPUR*** through my trainer-in-charge Mr. Chintoo Kumar.

PROBLEM STATEMENT

Imagine you are working as a data scientist in an insurance company in the USA. The company has collected the data of earlier travel insurance buyers. In this season of vacation, the company wants to know which person will claim their travel insurance and who will not. The company has chosen you to apply your Machine Learning knowledge and provide them with a model that achieves this vision.

OBJECTIVE

You are responsible for building a machine learning model for the insurance company to predict if the insurance buyer will claim their travel insurance or not.

For making the predictions use the test dataset provided.

BACKGROUND

1. Binary Classification

In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc. In our case, we have to classify the status of claim (if yes then 1 else 0).

1. Random Forest Classifier

The Random Forest is basically a supervised learning algorithm. This can be used for regression and classification tasks both. Random forest is one of the most used algorithms because of its simplicity and stability.

As the name suggests, "*Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset*." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Important points about Random Forest Classifier:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

1. Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems.

Important points about Logistic Regression:

* In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
* There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.
* We must include meaningful variables in our model.
* We should choose a large sample size for logistic regression.

SYNOPSIS

In this project I first observed and analyzed the dataset which was provided to me. After analysis, I discovered that the target label is the ‘Claim’ whose value can either be true (1) or false (0). So, I decided to solve this problem with binary classification problem solving approach. I processed the data through label binarizer and other logics (I created some functions) so that all the independent and dependent variables were in mathematically analyzable form.

Then I used two approaches; first, Random Forest Classifier and second, Logistic Regression to make predictions for the target label i.e., ‘Claim’. I created models by using these two approaches.

Then at last, I analyzed the models that I created for making predictions by using accuracy score, classification report and confusion matrix.

Then I imported the testing dataset provided to me and made the predictions.

CONCLUSION

So, through this project, I have successfully made predictions that if the insurance buyer will claim their travel insurance or not using Random Forest Classifier and Logistic Regression machine learning models.

REFERENCES AND BIBLIOGRAPHY

1. Notebooks provided by Chintoo Sir
2. An Introduction to Machine Learning by Miroslav Kubat (provided by Chintoo Sir)
3. https://scikitlearn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html?highlight=train\_test\_split#sklearn.model\_selection.train\_test\_split