Baseline Model for Insurance Fraud Detection Using Machine Learning

By Unknown unkown

WORD COUNT 652
TIME SUBMITTED 03-APR-2023 02:32PM
PAPER ID 98322770

Baseline Model for Insurance Fraud Detection Using Machine Learning

<student name>

<University name>

<Course Name>

<Professor Name>

Introduction

As Discussed in my previous report, the goal of this project, or work is to identify a machine learning model which achieves the best results out of a set of other machine learning models. The goal of this report, or first deliverable is to provide a baseline model that could be used as a reference to understand how well my other models are performing.

The task at hand is a classification task, given a particular sample of dataset, the goal is to predict if the insurance claim made is fraudulent. Hence my target variable within the dataset is: "fraud_reported."

Data Preparation

The dataset that I've chosen from Kaggle, contains close to 1000 samples within it, and has 39 features. Out of these a total of 19 columns are categorical in nature! In order fit this to any machine learning model we know that we have to perform some data pre-processing, hence I decided to use OneHotEncoder provided within the Sklearn package to encode these features and the resulting set has 1145 features.

This data is also split into 70% - 30% with 705 of the data being used as training set, and rest as Test set.

Baseline Model and Proposed Methodology

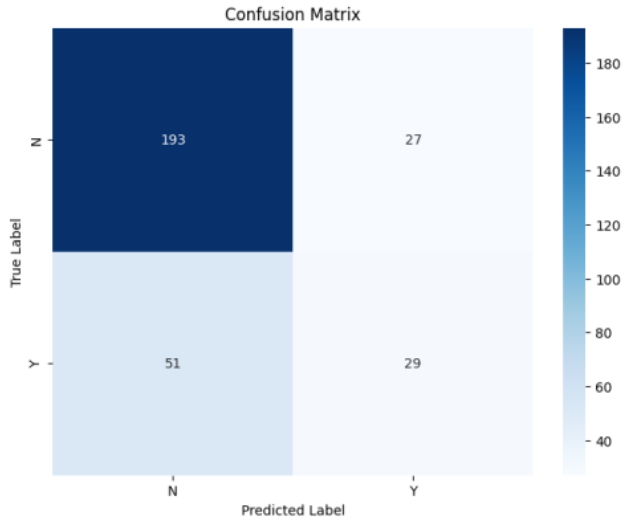
The model that I decided to as a reference/baseline model for this task, is Logisitic Regression. Primary reason being the fact that Logistic Regression is a very simple yet powerful classifier that is used in many classification tasks! Upon reading about it I realized that a lot of modern day NLP classification tasks are carried out well by Logistic Regression. This gave me some confidence to choose Logistic Regression as my baseline mode.

I trained a new instance of the Logistic Regression model on transformed data (after encoding categorical data). With no params, and default threshold I obtained the following results:

Default Threshold			
Accuracy	**Precision**	**Recall**	**F1**
0.74	**0.72**	**0.74**	**0.72**

Classification Metrics: Weighted Avg scores. For default Threshold.

I wanted to understand the classifications made, hence here is a confusion matrix for these early predictions:



Confusion matrix: Default Threshold.

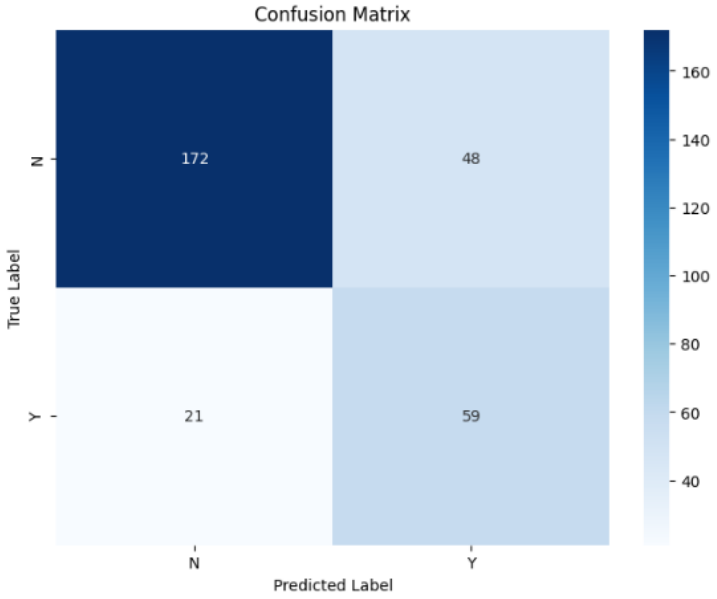
I personally consider these scores as average scores and not the best or optimal scores. In order further improve these metrics by small margin, I predicted probabilities, and then used a manually lowered threshold to get a better classification.

The following are the metrics that I achieved with a lower threshold:

Threshold: 0.2			
Accuracy	**Precision**	**Recall**	**F1**
0.78	**0.79**	**0.78**	**0.78**

Classification Metrics: Weighted Avg scores. for lower Threshold = 0.2.

The scores have slightly improved. Here is a confusion matrix for the new set of predictions made with threshold being 0.2



Confusion Matrix: Threshold = 0.2

The primary difference we can clearly see is that by lowering this threshold, we make our model more sensitive i.e. it will capture more actual fraud claims made. The trade off is the fact that it might produce more than usual false positives. But I consider this as a worthy trade off, if we think about it, no model is perfect and having a model that will capture any true positive is more important in this scenario than having a higher false positives.

Although these are not ground breaking results, I believe that these metrics could be a point of good reference, for training other models that I had mentioned in my initial report/proposal.

In order to outperform my baseline model I will try and perform some feature engineering, there are 39 features, surely some of these could be used for yielding a better understanding of the data, and search for optimal hyper params for my other models. And experiment with classification threshold/decision threshold to optimize each model to try and make it perform better.

References

ARPAN. (2022). Insurance Fraud Detection [Dataset]. Indian Institute of Management Calcutta & Kaggle. Retrieved from Insurance Fraud Detection | Kaggle

Baseline Model for Insurance Fraud Detection Using Machine Learning

ORIGINALITY REPORT

1%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.coursehero.com Internet	8 words — 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	OFF	EXCLUDE MATCHES	OFF