Machine Learning Nanodegree Capstone Project

Credit Card Fraud Detection using Supervised learning

Proposed by : Ahmed Roshdy

I. Definition

Project overview:

In this project I will build a credit card fraud detection solution for the credit card companies which shall be able to recognize fraudulent credit cards transactions .

I will start with the simple model architecture first before training and evaluating it. Then splitting the data , fit the model , and test the model by using predict function

Problem Statement :

The main objective of this project is to detect the fraud credit cards ,Credit card companies shall be able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase

Metrics:

The evaluation metric for this problem is simply precision, recall , and, f1-score

II. Analysis

Data Exploration:

The dataset used in this project is produced by kaggle website and available to download from https://www.kaggle.com/mlg-ulb/creditcardfraud

the dataset is labeled and consisting of 284807 rows and 31 columns , and our target feature that we should predict is the Class column , and if the value of the class is 0 that means it’s transaction without fraud , and if it is equal 1 that means it’s transaction with fraud.

Exploratory Visualization:

our target feature that we should predict is the Class column

Time	V1	V2	……	Amount	Class
0	0.0	-1.359807		149.62	0
1	0.0	1.191857		2.69	0
2	1.0	-1.358354		378.66	0
3	1.0	-0.966272		123.50	0
4	2.0	-1.158233		69.99	0

That’s sample of the column and sample of the rows

Solution statement , algorithms and technique:

The proposed solution to this problem is to apply supervised learning algorithm to detect if the new transaction is fraud or not , and we will use the logistic regression model , after a lot of trying I found that the logistic regression model give the best result.

Benchmark Model :

For the benchmark model, we will use logistic regression model.

And The 0 class (transactions without fraud) is predicted with 100% precision and recall whereas the 1 class (transactions which are fraudulent) has 90% precision. This means that 10% of the transactions which are fraudulent remain undetected by the system. This can be further improved by providing more training data.

III. Methodology

Data Preprocessing

We divided the data into a training set and validation set with an 80% training and 20% testing

Implementation

First we build a correlation matrix to see if there is any strong correlation between different variables in our dataset and to see how strong our linear relationships.

we trained the our logistic regression model and predict using the model and using classification_report function te check the precision , recall , and f1-score :

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56859
           1       0.90      0.68      0.77       103

    accuracy                           1.00     56962
   macro avg       0.95      0.84      0.89     56962
weighted avg       1.00      1.00      1.00     56962
```

Refinement :

For refinement I used GridSearchCV model and GradientBoostingRegressor for improving the result and these helped me to improve the results and I used more paramaters for the GridSearchCV to get the best_estimator_ , best_score_ , and best_params_ that can be using to get the best results ,

- And The best score across ALL searched params is: 0.9999618264982186

- The best parameters across ALL searched params: {'learning_rate': 0.03, 'max_depth': 6, 'n_estimators': 1000, 'subsample': 0.2}

IV. Results

Model Evaluation and Validation

I will start with the simple model architecture first, before training and evaluating it. Then splitting the data , fit the model , and test the model by using predict function,

At the end I used the grid search model to test the acuuracy and I got the best paramaters that I can use to get the best result ,

It was very useful for me to make refinment , and to evaluate my model and know if I got the best result or not.

Justification

Let me tell you that the final results found stronger than the benchmark result reported earlier and as I analyzed the result and the grid search cv model in last section , it was amazing , and it gave me the best result , and it was the best solution for solving this problem.

V. Conclusion

Free-Form Visualization

This is the result from the last solution which is :

```
Results from Grid Search
==================================================
======

 The best estimator across ALL searched params:
```

```
 GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse',
init=None,
                          learning_rate=0.03, loss='ls', max_depth=6,
                          max_features=None, max_leaf_nodes=None,
                          min_impurity_decrease=0.0,
min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0,
n_estimators=1000,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=0.2, tol=0.0001,
                          validation_fraction=0.1, verbose=0,
warm_start=False)

 The best score across ALL searched params:
 0.9999618264982186

 The best parameters across ALL searched params:
 {'learning_rate': 0.03, 'max_depth': 6, 'n_estimators': 1000,
'subsample': 0.2}
```

And the result from logistic regression model is :

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56859
           1       0.90      0.68      0.77       103

    accuracy                           1.00     56962
   macro avg       0.95      0.84      0.89     56962
weighted avg       1.00      1.00      1.00     56962
```

and this mean It's a good job !!!
The 0 class (transactions without fraud) is predicted with 100% precision and recall whereas the 1 class (transactions which are fraudulent) has 90% precision. This means that 10% of the transactions which are fraudulent remain undetected by the system. This can be further improved by providing more training data.

And that’s sample of the dataset :

Time	V1	V2	......	Amount	Class
0	0.0	-1.359807		149.62	0
1	0.0	1.191857		2.69	0
2	1.0	-1.358354		378.66	0
3	1.0	-0.966272		123.50	0
4	2.0	-1.158233		69.99	0

That’s sample of the column and sample of the rows

Reflection

You can check the earlier you will notify that I summarized the entire process that i used for this project

I think there weren’t *any* interesting aspects of the project

I think also that there weren’t any difficult aspects of the project

The final model and solution fit my expectations for the problem

Improvement

I just think the further improvements that if we just increase the training data in the future, we will get better results.

Of course there are many algorithms or techniques I researched that I did not know how to implement, but I would consider using if you knew how.