Credit card fraud detection using Decision trees

Saloni Goyal 2017ucp1061

Computer networks - Malaviya national institute of technology

0.1 Introduction

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.My project uses a machine learning model - Decision tree model to find fraudulent credit card transactions.

Decision trees are Machine learning algorithms that progressively divide data sets into smaller data groups based on a descriptive feature, until they reach sets that are small enough to be described by some label.

Main challenges involved in credit card fraud detection

- Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.
- Imbalanced Data i.e most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones
- Data availability as the data is mostly private.
- Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.

0.2 Dataset

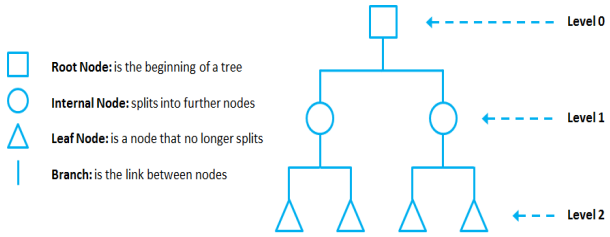
Dataset used in the project is "creditcard.csv" which is taken from website kaggle.com . Here's the link for downloading the dataset `https://www.kaggle.com/mlg-ulb/creditcardfraud`
The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
It contains only numerical input variables which are the result of a PCA transformation.Unfortunately, due to confidentiality issues, the original features and more background information about the data is not available. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

0.3 Decision tree classification

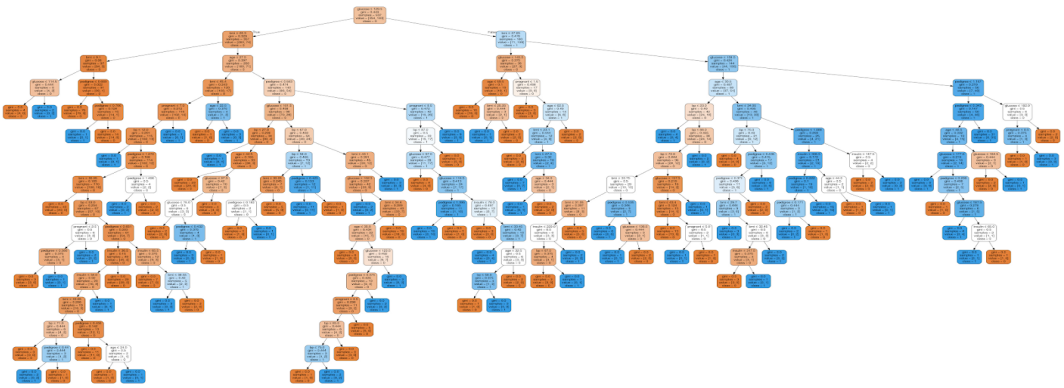
Decision trees are ML algorithms that progressively divide data sets into smaller data groups based on a descriptive feature, until they reach sets that are small enough to be described by some label.

DTs algorithms are perfect to solve classification (where machines sort data into classes, like whether an email is spam or not) problems.
DTs are also used to improve financial fraud detection. The MIT showed that it could significantly improve the performance of alternative ML models by using DTs that were trained with several sources of raw data to find patterns of transactions and credit cards that match cases of fraud.

DTs are composed of nodes, branches and leafs. Each node represents an attribute (or feature), each branch represents a rule (or decision), and each leaf represents an outcome. The depth of a Tree is defined by the number of levels, not including the root node.



When smaller data groups are generated using DT , all points having a common data group are assigned common output value ($\hat{y}$).

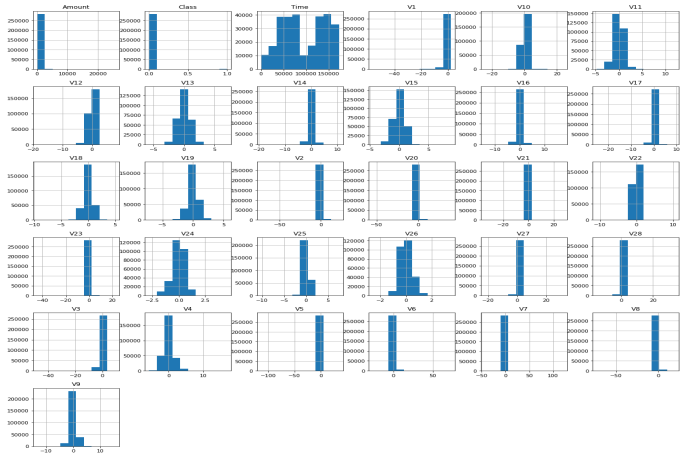


Above diagram shows how DT's form different data groups using different conditions.

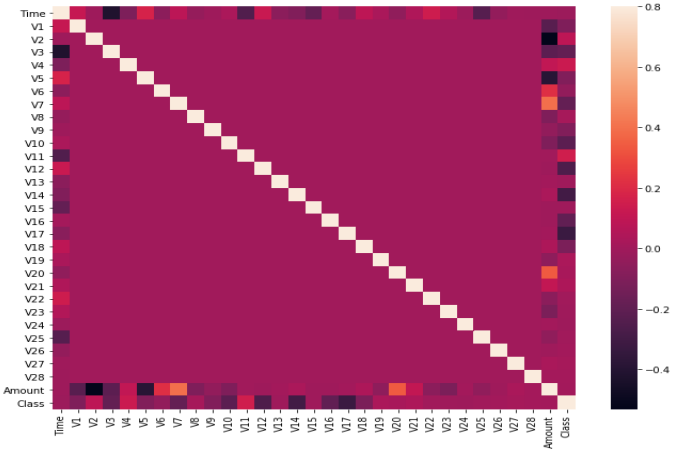
0.4 Project

- Dataset creditcard.csv contains 284807 total entries which is huge for our model , thus we will take only 10% (i.e 28481) of the given entries for our model.

- Properties of dataset attributes are shown with the help of histograms, these histograms shows mean,max,min etc properties of attributes.



- A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

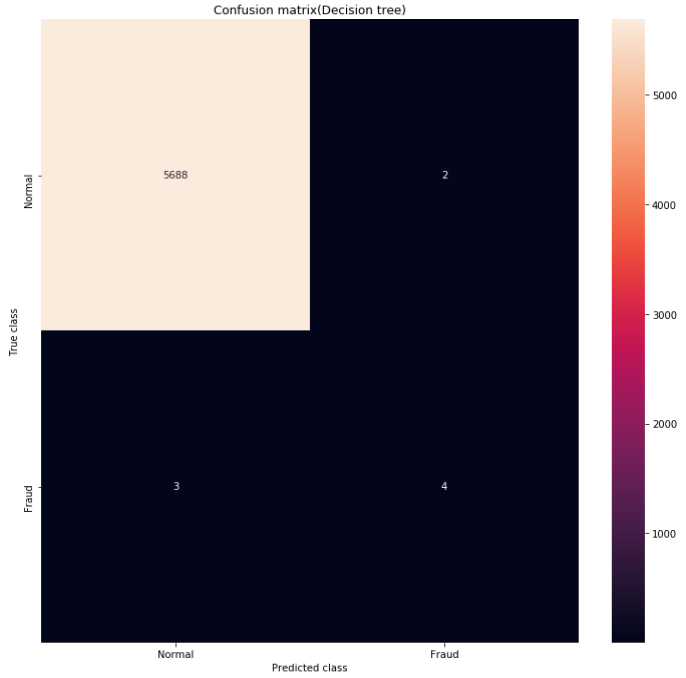


eg. in the given figure relation between amount and V2 is bad relationship ,amount and V7 is good relationship.

- Then we split the dataset into training set and test set. We will train our Decision tree model for training set and once trained we will see how much

accurately our model works for test set.I have chosen 20% of data from dataset for test set.

- At last after prediction the results for test set , we compare how many predictions we have got right and how many wrong using confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



We can see here that our model produced 5 wrong predictions and 5692 correct predictions

0.5 Metrics

0.5.1 Accuracy score

Accuracy is ratio of correctly predicted observation to the total observations.

$$AS = \frac{TP + TN}{TP + TN + FP + FN}$$

where: AS is accuracy score , TP is true positives , TN is true negatives ,FP is false positives , FN is false negatives

0.5.2 Precision score

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$PS = \frac{TP}{TP + FP}$$

where: PS is precision score , TP is true positives ,FP is false positives

0.5.3 Recall(sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$R = \frac{TP}{TP + FN}$$

where: R is recall , TP is true positives ,FN is false negatives

0.5.4 F1 Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution

$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision}$$

0.5.5 Matthews Correlation Coefficient(MCC)

The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of binary (two-class) classifications.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where: MCC is Matthews Correlation Coefficient , TP is true positives ,TN is true negatives, FN is false negatives, FP is false positives

0.6 Project results

S.no	Metric	Result
1	Accuracy score	0.9991%
2	Precision score	0.6667%
3	Recall	0.5714
4	F1 score	0.6154
5	Matthews correlation coefficient	0.6167

0.7 Conclusion

In this project we developed a credit card fraud detection system using Decision tree classifier which divide data sets into smaller data groups based on a descriptive feature, until they reach sets that are small enough to be described by some label having accuracy 0.9991%.

0.8 References

Jiang, Changjun et al. "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." IEEE Internet of Things Journal 5 (2018): 3637-3647.
https://www.sciencedirect.com/science/article/pii/S187705092030065X
https://towardsdatascience.com/the-complete-guide-to-decision-trees-28a4e3c7be14
https://www.kaggle.com