University of London

MSc Data Science and Artificial Intelligence

Big Data Analysis- Coursework 2

Classification of Credit card Transactions

Student ID- 210590034

September 2022

Contents

Introduction

Big data solutions are highly significant challenges that are faced by several organizations. There are various methods, solution options and software that must be put into consideration in order to guarantee an effective outcome of big data solution implementation. As traditional database management systems are not sufficient anymore, Big Data solutions helps organizations unlock the strategic values and take full advantage of the vast amount of data produce, examples of the most popular Big Data solutions include Apache Hadoop and Apache Spark. Hadoop is a distributed file system that allows for the processing of data sets, making it one of the leading software used due to its ability to store large amounts of data on commodity hardware. Spark is a key tool in data processing as it is a framework that sits on top of Hadoop's HDF, known for its simplicity, flexibility, and the support that can be found from the community.

This project aims to explore the benefit of using both HDFS and Spark for analyzing big amounts of data and generating valuable insights. The data set used for this project was obtained from Kaggle, which is an online community that contains a vast collection of opensource datasets. The data is simulated credit card transactions, with legitimate and fraudulent transactions starting from the 1st of January 2019 until the 31st of December 2020.The size of data used is around 478MB. Which might be on the smaller side considering the industry it is related to and the potential of HDFS and Spark. However, the methods used should be sufficient regardless of how big or small the data used is.

Data

The data set intended to be used in this analysis is divided into two files, fraudTrain.csv and fraudTest.csv. They both contain 23 columns with 1296675 rows in fraudTrain.csv and 555719 rows in fraudTest.csv.The 23 columns are as mentioned in the table below:

Table 1 Brief description of each of the 23 columns used in the data set. This Data has 0 duplicated values and 0 missing values.

Trans_date_trans_time	Date and time of transaction. In yyyy-dd-mm hh:mm:ss format
cc_num	Credit card number associated with the transaction. Column has 999 distinct values
Index	Unique identifier for each row
merchant	Merchant name. Has 693 distinct values
category	Category of merchant. Has 14 distinct values
amt	Amount of transaction. Min value = 1, max value = 28,948.9, mean = 70, median = 47.5
first	First name of credit card holder
last	Last name of credit card holder
gender	Gender of credit card holder. 54.8% were female, while 45.2% were male
street	Street address of credit card holder
city	City of credit card holder. Has 906 distinct values
state	State of credit card colder. Has 51 distinct values
zip	Zip of credit card holder
lat	Latitude location of credit card holder. Has 983 distinct values
long	Longitude location of credit card holder. Has 983 distinct values
city_pop	Credit card holder's city population. Min value = 12, max value = 2,906,700, median = 2,443
job	Job of credit card holder. Has 497 distinct values
dob	Date of birth of credit card holder. Has 984 distinct values
trans_num	Transaction number. Has 1,852,394 distinct values
unix_time	UNIX time of transaction
merch_lat	Latitude location of merchant
merch_long	Longitude location of merchant
is_fraud	Fraud Flag. Target Class

Hypothesis

Generating machine learning algorithms can assist in the identification and prevention of fraudulent credit card transactions.

Aims/Objectives

The main objective of this coursework is to utilize a distributed computing framework such as PySpark to preform big data analysis. This can then benefit us in trying to analyze credit card transactions to find variables that have a correlation with fraud cases. Finally, this coursework aims to build a classification model that would flag credit card transactions that are predicted to be fraudulent.

Methodology

Data source

The big data set that will be used in this coursework was retrieved from Kaggle (Link in references) which is an online platform that provides large data sets to be used in analysis studies.

Moving data to HDFS

Files will be moved from local PC to Lena using command in terminal:

```
scp fraudTrain.csv {username}@lena.doc.gold.ac.uk:./Big_Data/CW2
scp fraudTest.csv {username}@lena.doc.gold.ac.uk:./Big_Data/CW2
```

Files will be moved from Lena to HDFS using command terminal:

```
hadoop fs -copyFromLocal /Big_Data/CW2/fraudTrain
hadoop fs -copyFromLocal /Big_Data/CW2/fraudTrain
```

Importing data into PySpark notebook from HDFS

Train data will be imported from HDFS using the following PySpark command:

```
df_train = sc.read.csv('fraudTrain.csv', header=True, inferSchema = True)
```

Train data will be imported from HDFS using the following PySpark command:

```
df_test = sc.read.csv('fraudTest.csv', header=True, inferSchema = True)
```

Data cleaning and preparation

Initially, data will be checked for duplicate rows and null values, then the date of birth and trans_date_trans_time columns will be used to calculate age of customer at the time of the transaction. Subsequently, the trans_date_trans_time column will be used to extract the data that would be beneficial to the analysis such as corresponding number of months, day of the month, and time of the day. In order to obtain the customers full name, the first and last columns will be concatenated. The following columns will be dropped as they will not be beneficial for this specific analysis: index, trans_date_trans_time, first, last, street, zip, lat, long, date of birth, trans_num, unix_time, merch_lat, merch_long

Exploratory Data Analysis (EDA) - Univariate and multivariate analysis

Univariate analysis will be carried out by checking balance of the is_fraud variable, which is the target variable, along with checking summary metrics for numerical variables. Furthermore, multivariate analysis will be done by examining the correlation between all the variables and the is_fraud variable. Creating bins for the numerical variables with large range to check the correlation with the is_fraud variable is another multivariate analysis that will be used.

Model Building and Validation

Index categorical columns that may have some correlation with the is_fraud column by assigning a numerical value to the column. However, only columns that are found to have some correlation with the target value will be used, a few examples of these include category, city, state, and job. Next, the following columns will be dropped: cc_num, merchant, category, gender, city, gender, state, city pop, job, full_name, amt_bucket, and amt_bin as they will be either indexed as a numerical column from the

previous step or will not have a significant correlation with the is_fraud column. All the remaining columns will then be converted into type integer, and a features vector will be created. The data will be split into train and test data sets and the CART model will be used to classify whether the transactions are legitimate or fraudulent. To validate the model, precision, accuracy, sensitivity, specificity, will be calculated for the CART model. Finally, another model to that will be used to classify legitimate versus fraudulent transactions is the Logistic Regression model. And similarly, precision, accuracy, sensitivity, and specificity will be calculated for the Logistic Regression model.

Conclusion

Ultimately, this coursework emphasizes the reason that PySpark is a preferred interface to use when dealing with big data. This was evident in the difference of data processing time when compared to Python. However, one of the main drawbacks witnessed while using PySpark instead of Python was that PySpark lacked the flexibility of Python for data manipulation. Additionally, PySpark lacked the functions found in Python as it requires multiple lines of code to achieve what could be done in Python using considerably less code.

In regard to the data, the columns that were chosen (amt, month, day, hour, age_at_transaction, category, city, state, and job) seemed to be good predictors of whether the transaction was fraudulent or not. Using these columns to build the CART model proved beneficial as the model was able identify over 50% of the fraudulent cases. However, the Logistic Regression model was not successful as it failed to identify any of the fraudulent cases.

References

- Lectures by Prof. Raju Chinthalapati
  - Topics 6, 7, 8, and 9

- Credit Card Transactions Fraud Detection Dataset | Kaggle

- 1 Introduction - Data Analysis with Python and PySpark (manning.com)

- Data Analysis With Pyspark Dataframe (nbshare.io)

- A Guide to exploit Random Forest Classifier in PySpark | by Manusha Priyanjalee | Towards Data Science

- Classification in PySpark | Chan`s Jupyter (goodboychan.github.io)

- Pyspark with Python - YouTube