**MINI-PROJECT REPORT**

**MACHINE LEARNING**

**IN BIG DATA USING PYSPARK IN HADOOP PLATFORM**

Submitted By:

Jeeva Mary Loui

&

Leno Mathew

Metro College of Technology

August-04 2019

**TABLE OF CONTENTS**

Linear Regression Model….……………………………………………………………….…3

Introduction…….………………………………….………………………………….….3

Implementation Details.…………………………………………….…………….…4

Results and Discussions.……………………………………………………………10

Recommendation………….………………………………………………………….10

Classification Model…….…………………………………………………………………….11

Introduction………………………………………………………………………………11

Implementation Details…………………………………………………………….12

Conclusion…………………………………………………………………………………20

# **Project part-1: Linear Regression Model: Google Travel Review Rating**

# **Introduction**

The dataset for the regression project was taken from the UCI database. The dataset contains 5456 data points based on Google reviews given to European tourist attractions in 24 categories. Google user rating ranges from 1 to 5 and average user rating per category is calculated.

The target value is the average ratings on monuments and how it gets affected by all other average ratings. As most tourists visit Europe to see the great monuments all other parameters will influence the entire experience of a tourist.

|  |  |
| --- | --- |
| Data Source | UCI Machine Learning Repository  <https://archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings> |
| Variables Type | Numerical |
| No. of Observations | 5456 |
| No. of Variables | 25 |
| File type | CSV |
| Analysis Type | Linear Regression |
| Data Set Characteristics | Multivariate |

**Project Objectives**

* To understand the factors that influences the overall experience of tourists who visit Europe and which factors are strongly correlated
* By analyzing strongly correlated fields government and private agencies can identify the key areas to invest so as to maximize customer experience and industry income

**Following attributes are used to rate the tourist attractions:**

1. Unique user id

2. Average ratings on churches

3. Average ratings on resorts

4. Average ratings on beaches

5. Average ratings on parks

6. Average ratings on theatres

7. Average ratings on museums

8. Average ratings on malls

9. Average ratings on zoo

10. Average ratings on restaurants

11. Average ratings on pubs/bars

12. Average ratings on local services

13. Average ratings on burger/pizza shops

14. Average ratings on hotels/other lodgings

15. Average ratings on juice bars

16. Average ratings on art galleries

17. Average ratings on dance clubs

18. Average ratings on swimming pools

19. Average ratings on gyms

20. Average ratings on bakeries

21. Average ratings on beauty & spas

22. Average ratings on cafes

23. Average ratings on view points

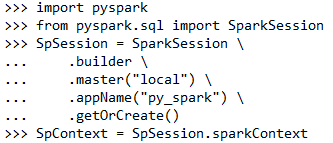
24. Average ratings on gardens

25. Average ratings on monuments

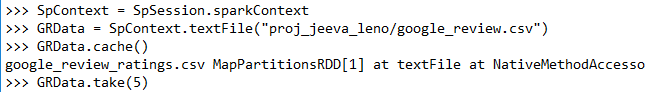
**Implementation Details**

**Exploratory Data Analysis for Linear Regression:**

**Step 1: Importing Libraries**



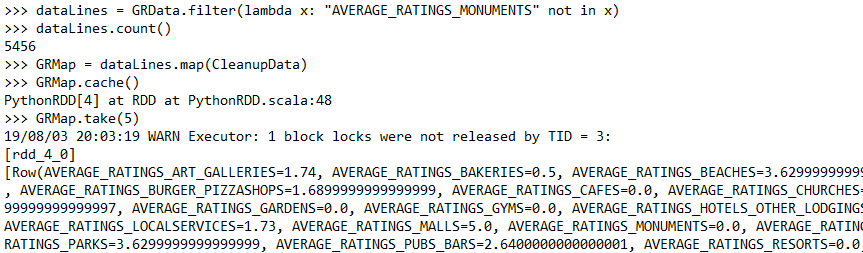
**Step 2: Importing the Dataset**



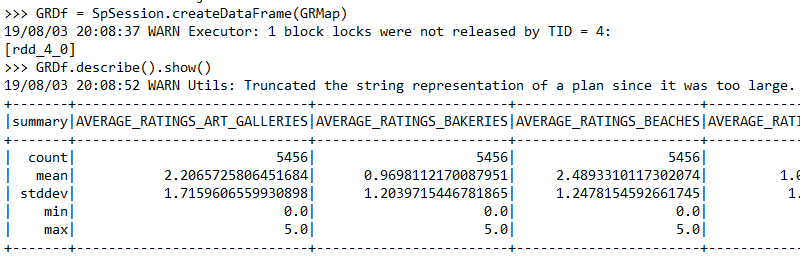
**Step 3: Data Cleanup**



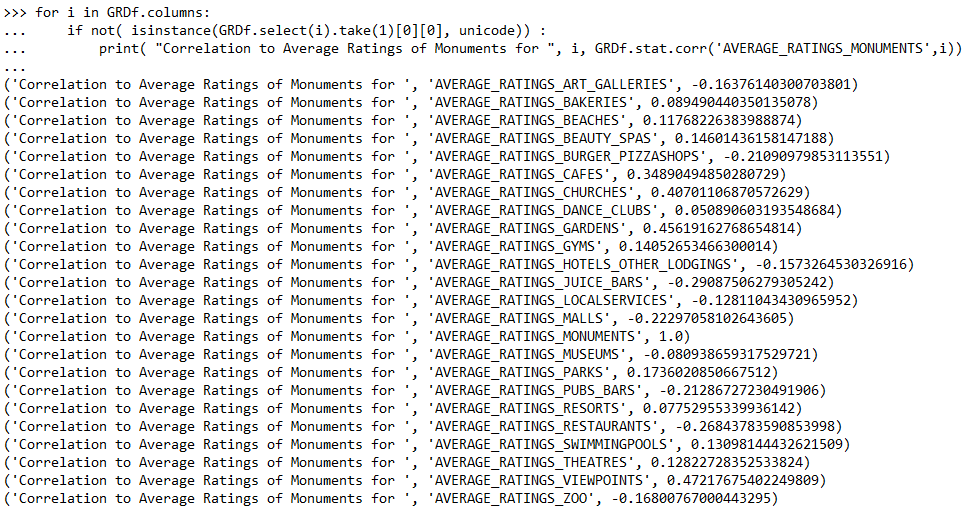
**Step 4: Run Map for cleanup**



**Step 5: Create a data frame**

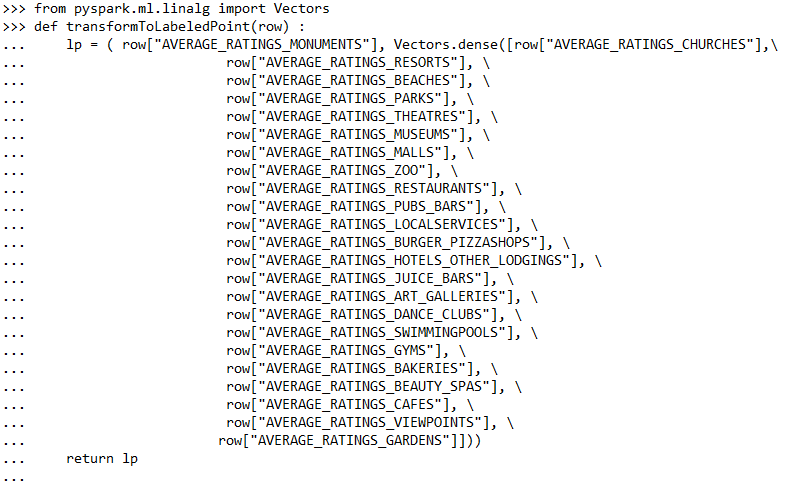


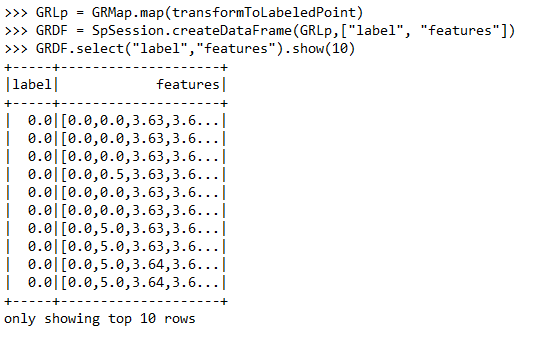
**Step 6: Find correlation between predictors and target**



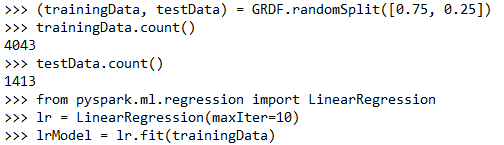
From the above results, its clear that all the independent variables are poorly correlated with the target variable. However, the most important correlations are average ratings on churches, cafes and view points and gardens

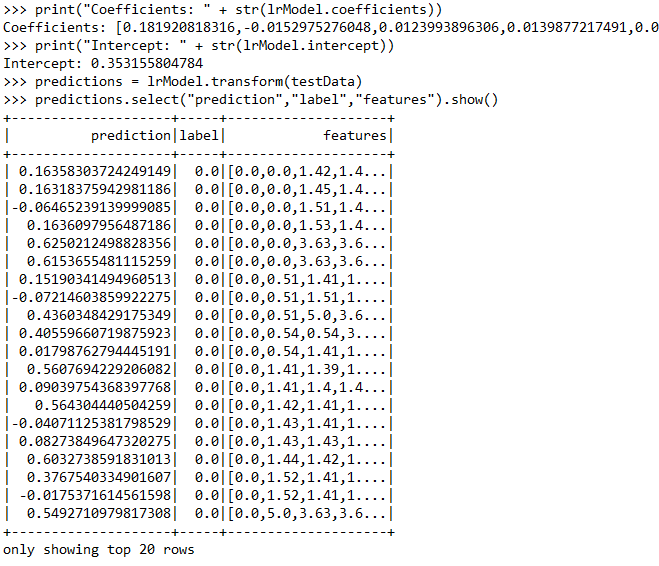
**Step 7: Transform to a Data Frame for input to Machine Learning**



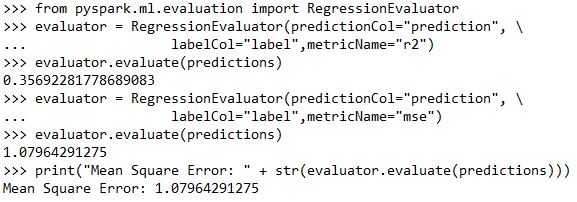


**Step 8: Perform Machine Learning**

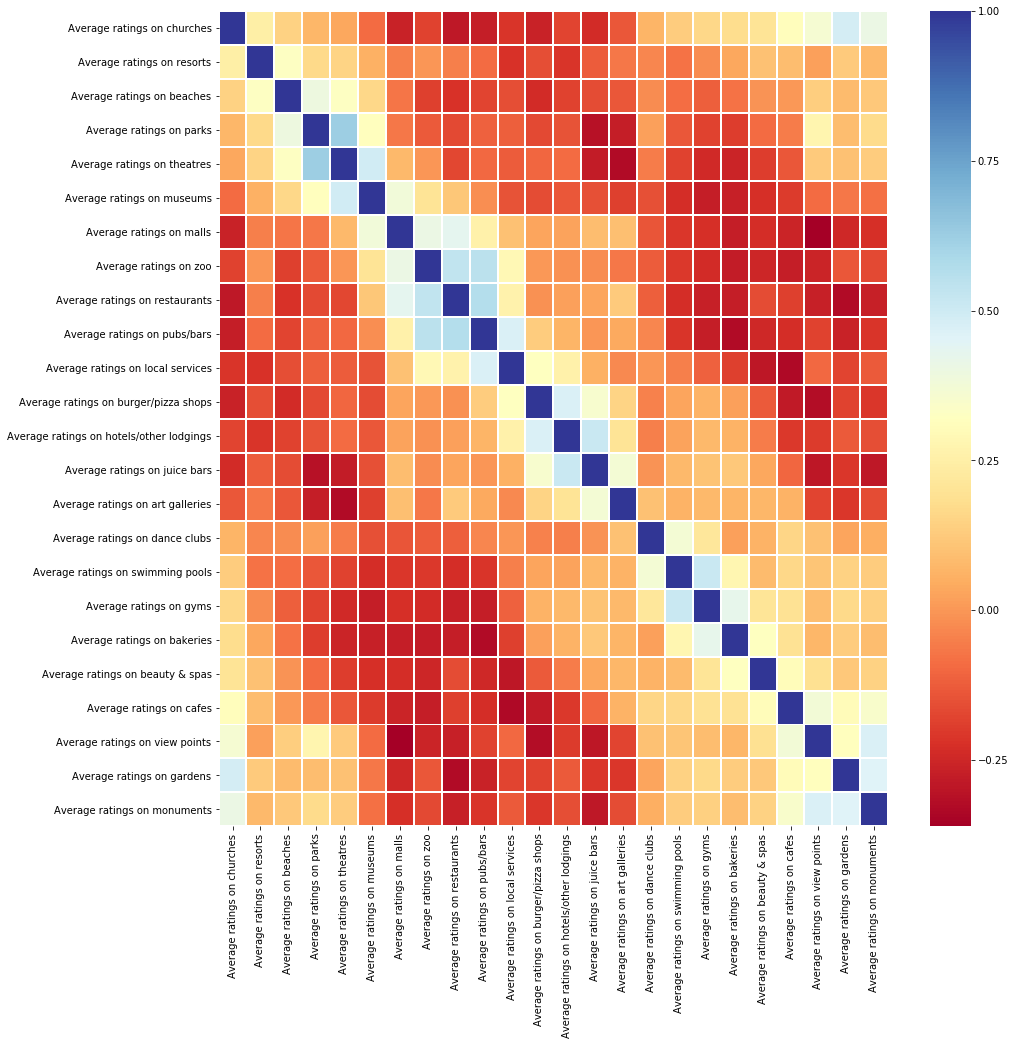




**Step 9: R2 Score and Mean Square Error**



**Step 25: Heat map for the data**



# **RESULTS AND DISCUSSIONS**

* The correlation between the independent variable and the target variable is very weak
* A positive experience in one of the attractions can improve the overall travel experience of the travellers
* R2 value is low – 0.36, indicating that the model won’t be of practical use
* Google Ratings of zero should be revaluated as ratings starts from 1

**RECOMMENDATION**

A direct influence of one parameter on the average rating of monuments seems to be low. Ratings of several attractions is given as zero while the minimum Google rating is 1, hence the data should be revaluated for better results

# **Project part -2: Build a Customer Churn Predictor Using Decision Tree Classification Method**

# **INTRODUCTION**

Customer churn, when a customer ends their relationship with a business, is one of the most basic factors in determining the revenue of a business. For a business it is important to know which of the customers are loyal and which are at risk of churning, and also its important to determine the factors that affect these decisions from a customer perspective. This code pattern explains how to build a machine learning model and use it to predict whether a customer is at risk of churning

|  |  |
| --- | --- |
| Data Source | Kaggle.com  <https://www.kaggle.com/zagarsuren/telecom-churn-ibm-watson> |
| No. of Observations | 7043 |
| No. of Variables | 22 |
| File type | CSV |
| Data Set Characteristics | Multivariate |

**Project Objectives**

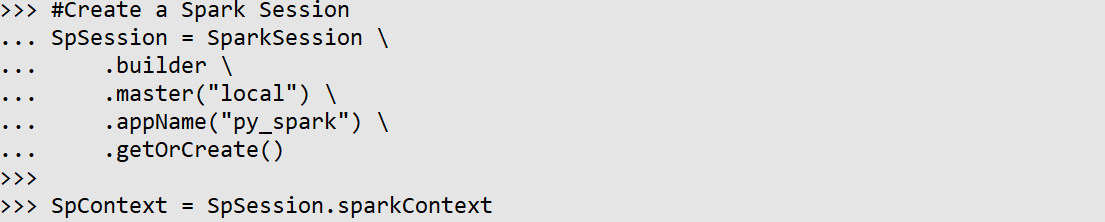
* Find the factors which are correlated to churn of customer in the telco company
* Run the decision tree classifier to build a model to predict the churn of customer.
* The report analyzes and gives findings about the Accuracy, F1 score,

**Following attributes are used in the data:**

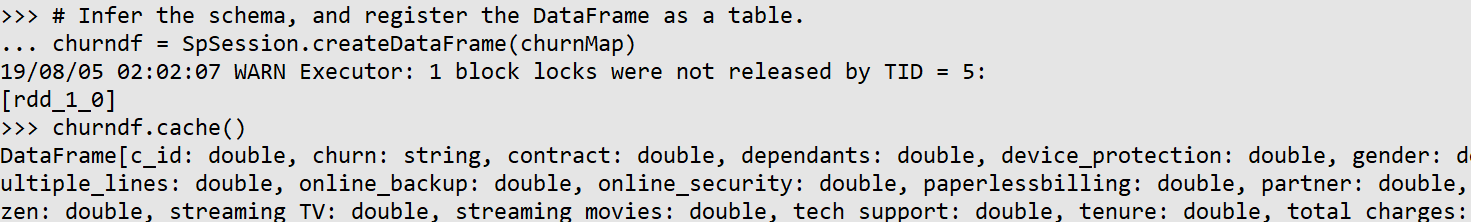
1. ‘Customer\_ID’ : Unique identification given to each customer.
2. 'gender': Sex of Customer.
3. 'SeniorCitizen': Age classification.
4. 'Partner': Whether customer has a partner in sharing connection or not.
5. 'Dependents': Whether customer has a dependant or not.
6. 'tenure'. Customer connection duration.
7. 'PhoneService': Whether customer use a phone service or not.
8. 'MultipleLines': Whether customer uses multiple lines for connection or not.
9. 'InternetService': Whether customer uses internet service or not.
10. 'OnlineSecurity': Whether customer uses online security or not.
11. 'OnlineBackup': Whether customer uses online backup or not.
12. 'DeviceProtection': Whether customer uses online backup facility or not.
13. 'TechSupport': Whether customer uses tech support or not.
14. 'StreamingTV': Whether customer stream TV or not.
15. 'StreamingMovies': Whether customer stream movies or not.
16. 'Contract': Whether customer is in a contract or not.
17. 'PaperlessBilling': Whether customer use paperless billing or not.
18. 'PaymentMethod':Payment mode of customer.
19. 'MonthlyCharges':Monthly charge of customer.
20. 'TotalCharges':Total Charges of Customer.
21. 'Churn':The target or Churn.

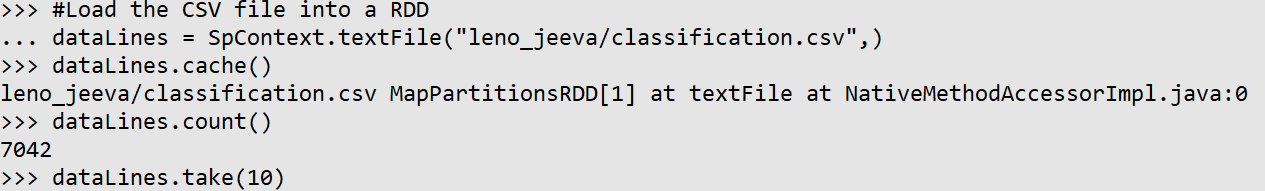
**Detailed Description of the Project:**

**Step 1: Create a Spark Session in Hadoop Platform**

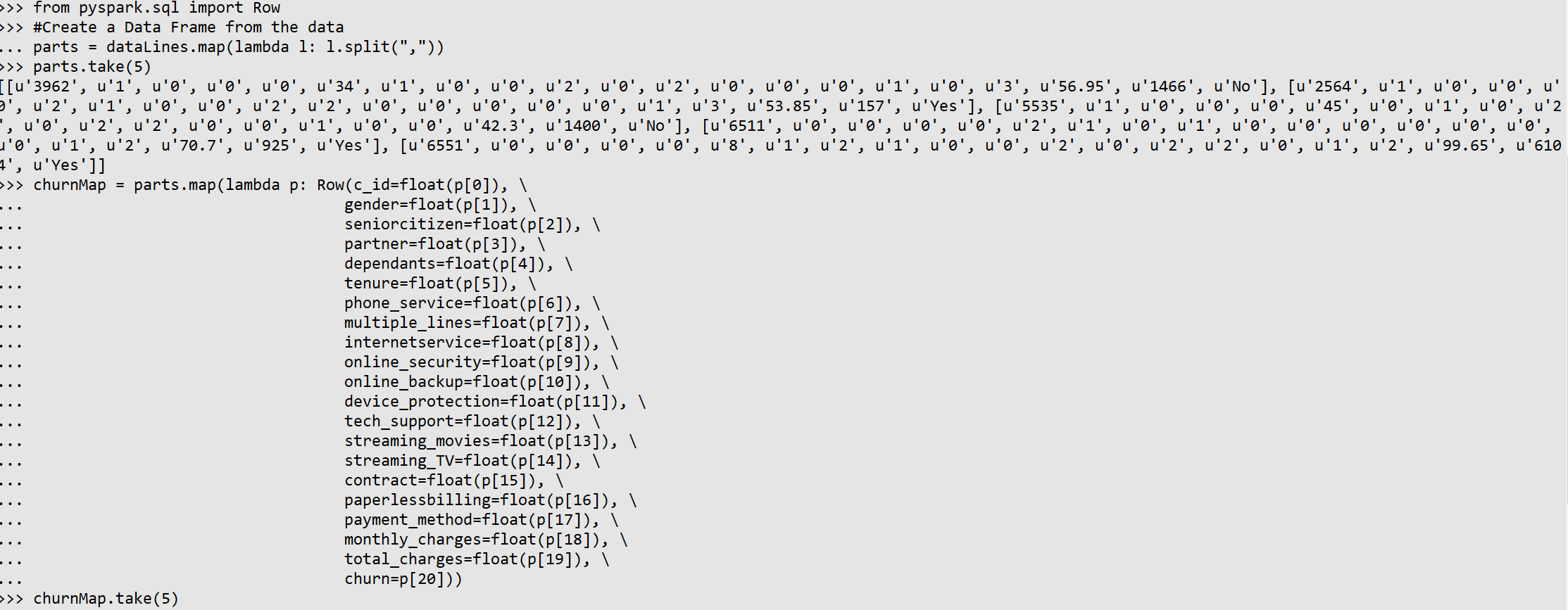


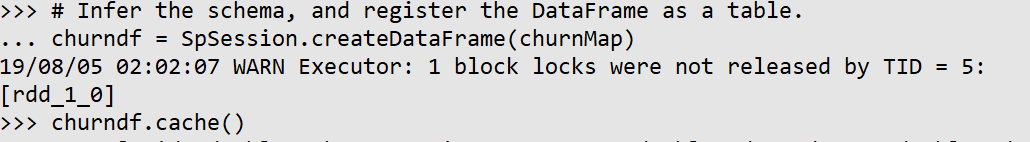
**Step 2: Data Analysis and Create RDD**

****

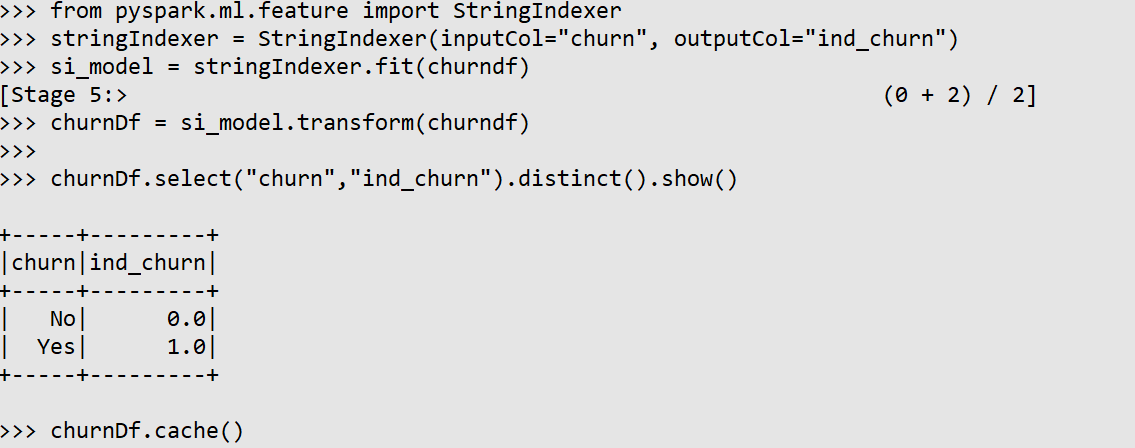
****

**Step3: Create Dataframe from RDD**

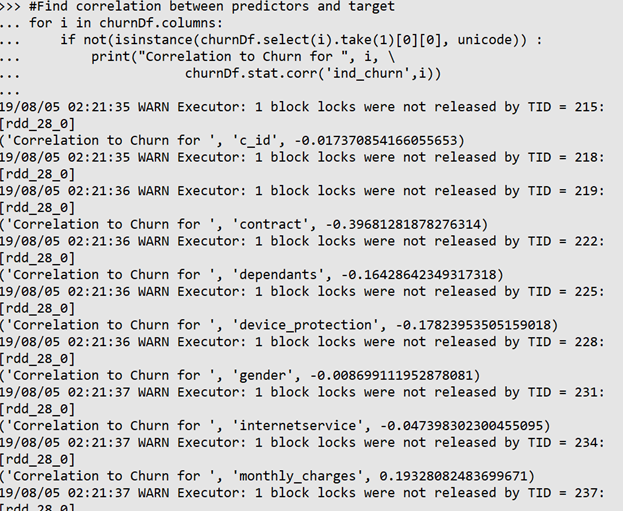


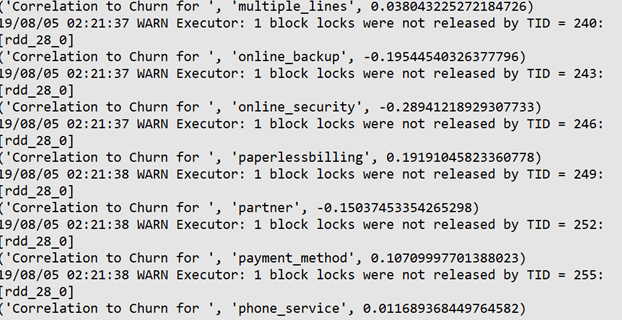


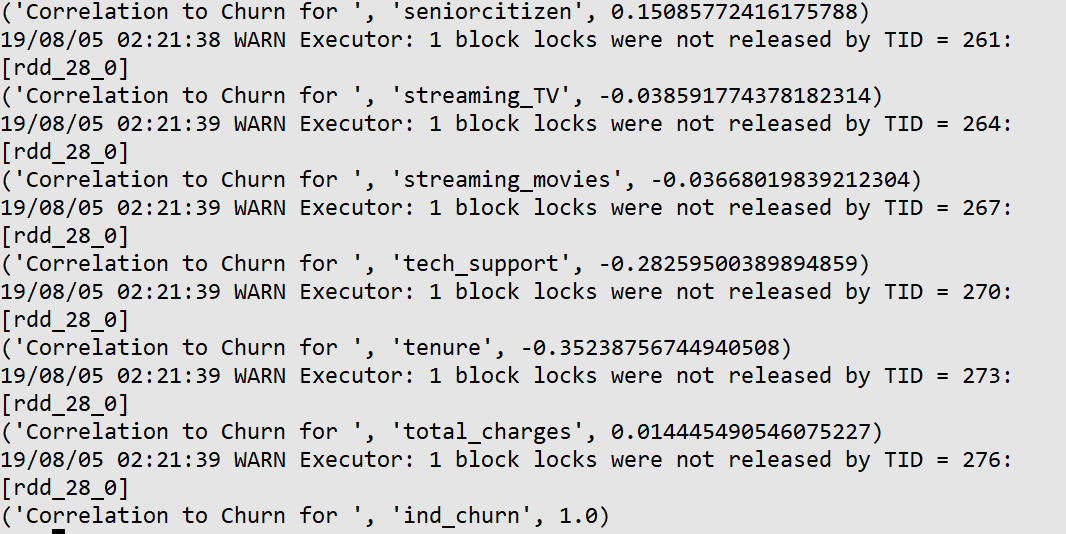
**Step 4: Label And Index the target using String Indexer**



**Step 5: Run Correlation Analysis to study the correlation of each variable to target**

****

****

****

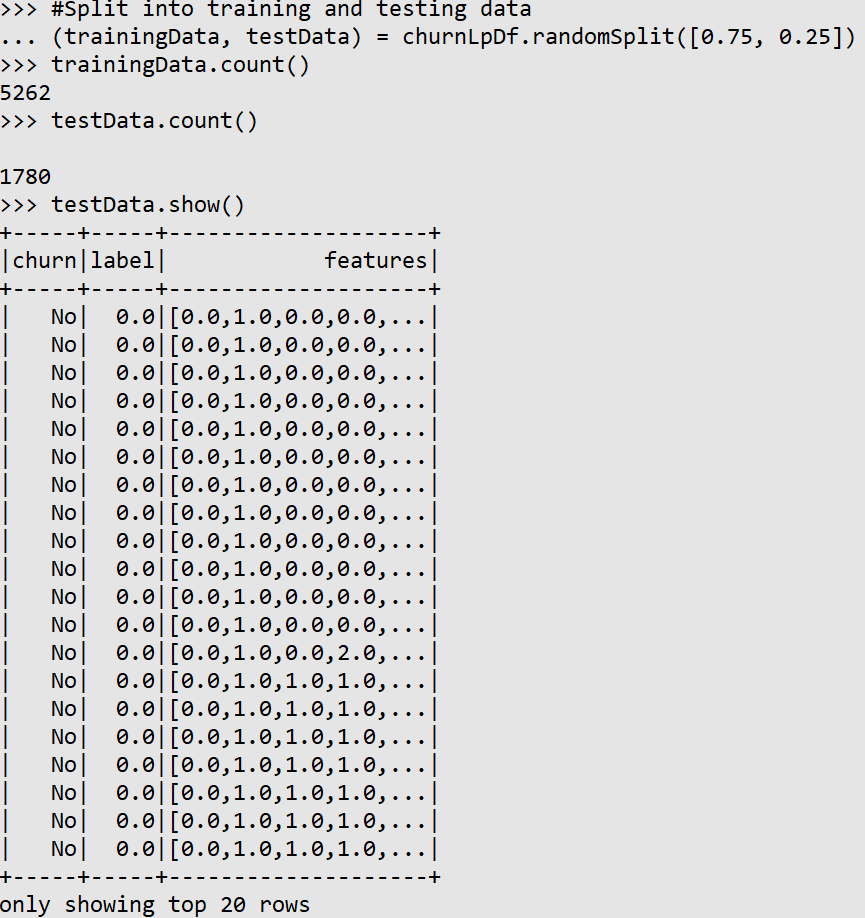
**Step 6: Select variable with strong correlation score and transform rows to labeled points.**

The features contract, tenure, tech\_support, online\_security, paperlessbilling, online\_backup, monthly\_charges, device\_protection have correlation with target churn higher than 0.17 and the rest of features are eliminated.

****

**Step 7: Split dataframe to train and test to run machine learning model.**

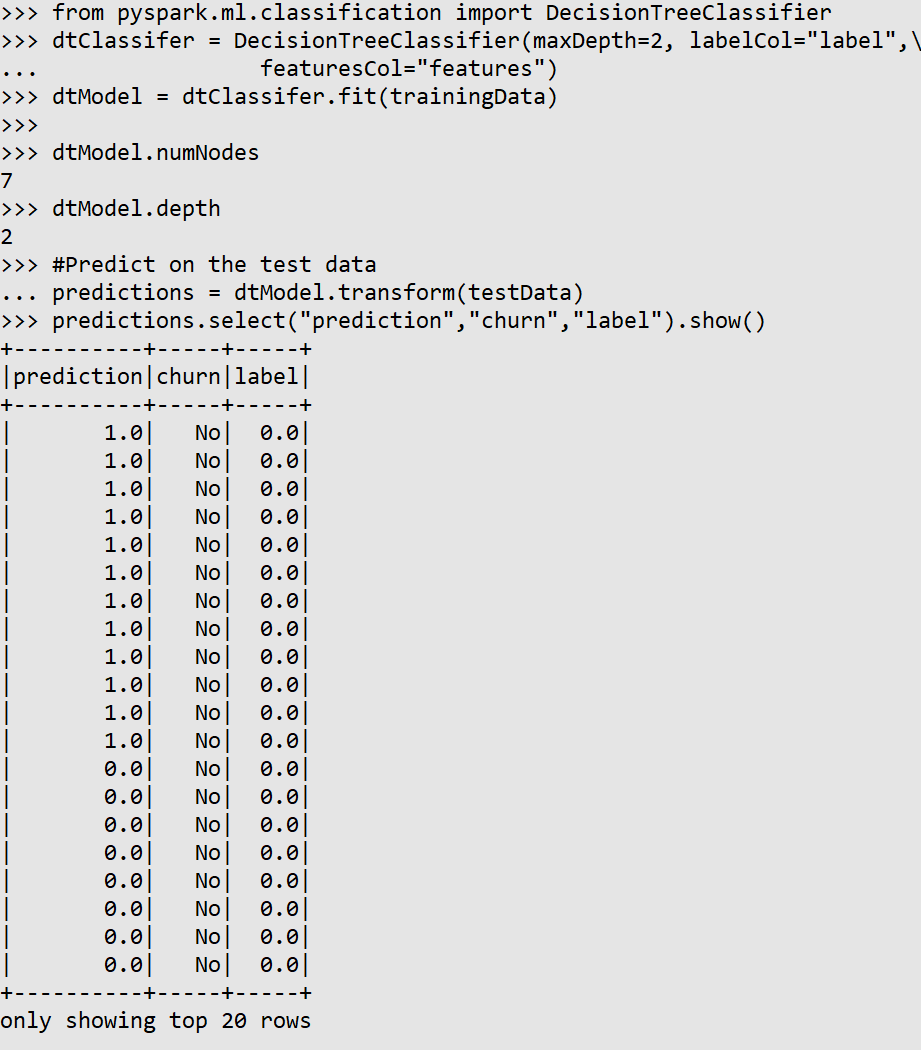
The dataframe is split to train and test using random split with a ratio of ¾ and ¼

****

**Step 8: Run the Decision Tree Classification Model**

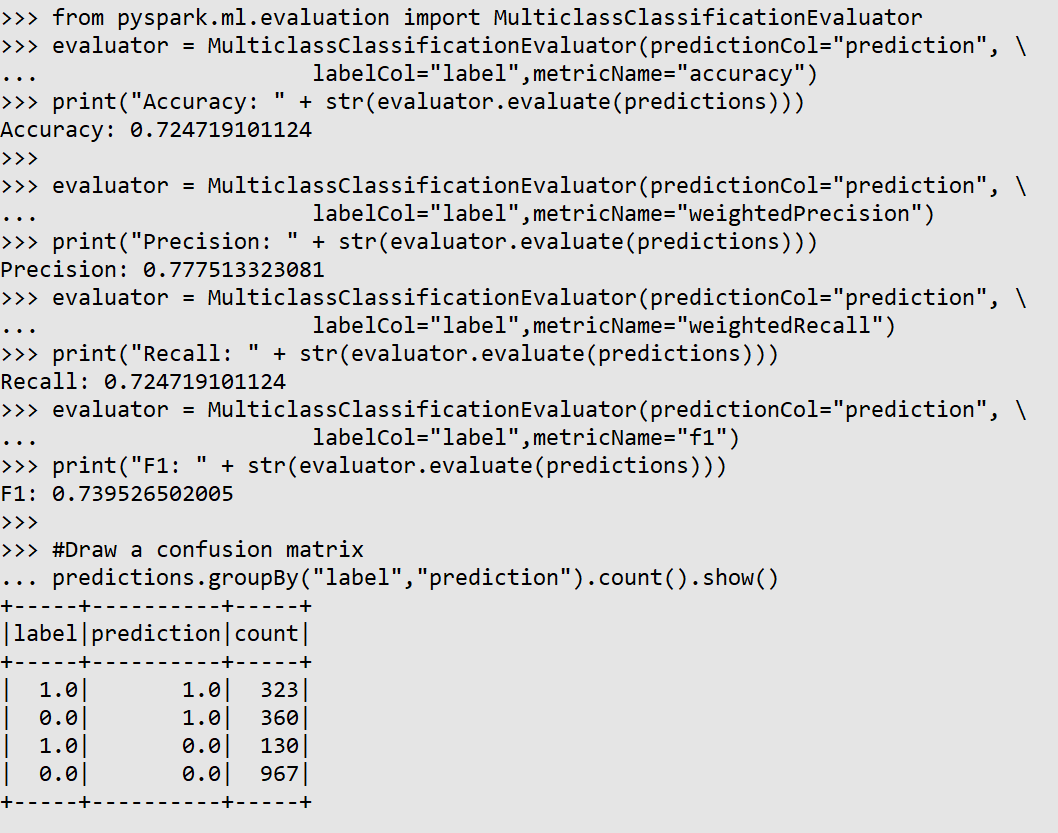
The decision tree classification model is build with the train set with number of nodes equal to seven and depth of Two.

The predictions are made with the trained model. And the predictions are displayed.



**Step 9: Classification Evaluation using metrics and create confusion matrix**

From the pyspark evaluation library MulticlassificationEvaluator is run to see how successful the predictor model is.



**CONCLUSION**

The model to predict the churn of customer using decision tree classifier is built successfully with an accuracy of .72, Precision of .77, Recall of .72 and F1 score of .72.

The scores say the model can be employed with 72% accuracy. And thus, churn of customer can be predicted with this model.