**BIA – 678 Big Data Technologies**

**Fall’ 21**

**Customer Churn Analysis of a Telecom Company**

**Team 5**

**Nidhi Chovatiya**

**Janki Patel**

**Malarvizhi Pandian**

**Mugdha Kamat**

Table of Contents

[**Introduction** 3](#_Toc90068680)

[**Business Understanding** 3](#_Toc90068681)

[**Project Goal** 3](#_Toc90068682)

[**Data Preparation** 5](#_Toc90068683)

[**Exploratory Data Analysis** 5](#_Toc90068684)

[**Attribute Selection** 7](#_Toc90068685)

[**Handling Missing Values** 7](#_Toc90068686)

[**Correlation Matrix** 8](#_Toc90068687)

[**Encoding String Variables** 8](#_Toc90068688)

[**Data Splitting** 8](#_Toc90068689)

[**Feature Scaling** 9](#_Toc90068690)

[**Over-sampling using SMOTE** 9](#_Toc90068691)

[**Vector Assembler** 9](#_Toc90068692)

[**Modeling** 10](#_Toc90068693)

[**Classification algorithms used** 10](#_Toc90068694)

[**Principal Component Analysis** 10](#_Toc90068695)

[**Evaluation** 11](#_Toc90068696)

[**Algorithm performance with PCA** 11](#_Toc90068697)

[**Impact of scale on quality of analysis** 11](#_Toc90068698)

[**Impact of scale on time** 12](#_Toc90068699)

[**Impact of Parallel Computation** 12](#_Toc90068700)

[**Conclusion** 13](#_Toc90068701)

[**Recommendation** 13](#_Toc90068702)

[**References** 14](#_Toc90068703)

[**Appendix** 14](#_Toc90068704)

# **Introduction**

With the fast expansion of the telecommunications sector, service providers' customer bases tend to grow. Keeping current consumers is a huge problem in a highly competitive industry. The process through which a customer discontinues using a company's products or services is known as Customer Churn. It is said that the expense of recruiting new clients is substantially higher than the cost of keeping existing customers. As a result, it is critical that telecommunications businesses employ sophisticated analytics to evaluate consumer behavior and subsequently anticipate if a client would quit the company. Churn is most typically stated as the percentage (or quantity) of service users that terminate their membership within a given time frame.

# **Business Understanding**

## **Project Goal**

The goal of this project is to predict if a customer is at a high risk of churning or not. Classifying the potential churning customer will assist the firm in taking all precautionary measures to retain the customer with the firm’s service. As the customers are classified into two groups - churn and non-churn, this project is considered as a binary classification problem. Expected outcomes from undertaking this project are as follows.

1. Report the firm on the factors that affect the churn rate to a higher extent, which in turn helps them to resolve the issue.
2. Utilize a few of the best classification algorithms to classify customers into two groups with the highest accuracy (90% and above) possible.

**Data Understanding**

**Data Description**

Dataset consists of customer data of a telecom company including information on customer location, call details, subscription details, income, and billing amounts. Each row denotes a unique customer, and the initial dataset contains data of 2154048 customers. The raw data contains 18 independent variables and one target variable - ‘CHURN’. Detailed data description as follows:

Table

Description automatically generated

Table : Dataset variables and description

Among the independent variables, there are 3 categorical variables - REGION, TENURE, TOP\_PACK. The other 15 independent variables are continuous variables. Target variable CHURN consists of two values ‘0’ denotes non-churn customer and ‘1’ denotes churn customer.

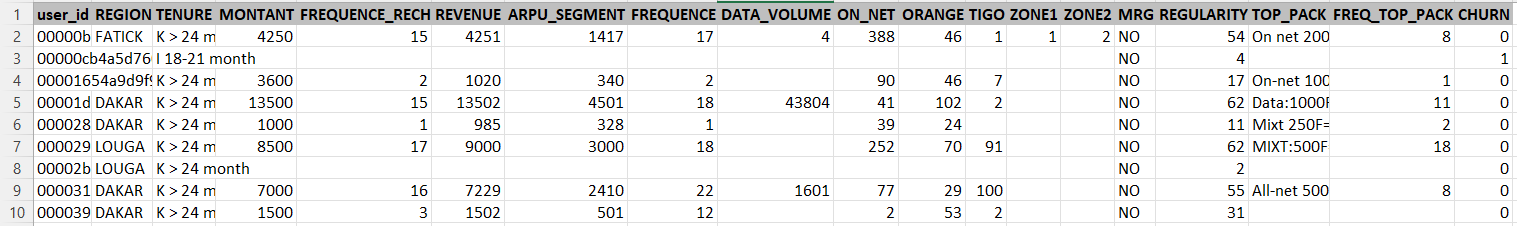


Figure : Sample raw data

# **Data Preparation**

## **Exploratory Data Analysis**

Chart, bar chart

Description automatically generatedExploratory Data Analysis (EDA) is performed on the dataset to understand the characteristics of each variable in the dataset. It is important to get these insights on the features before going into data preprocessing.

Figure : Percentage of missing values in each column of the dataset

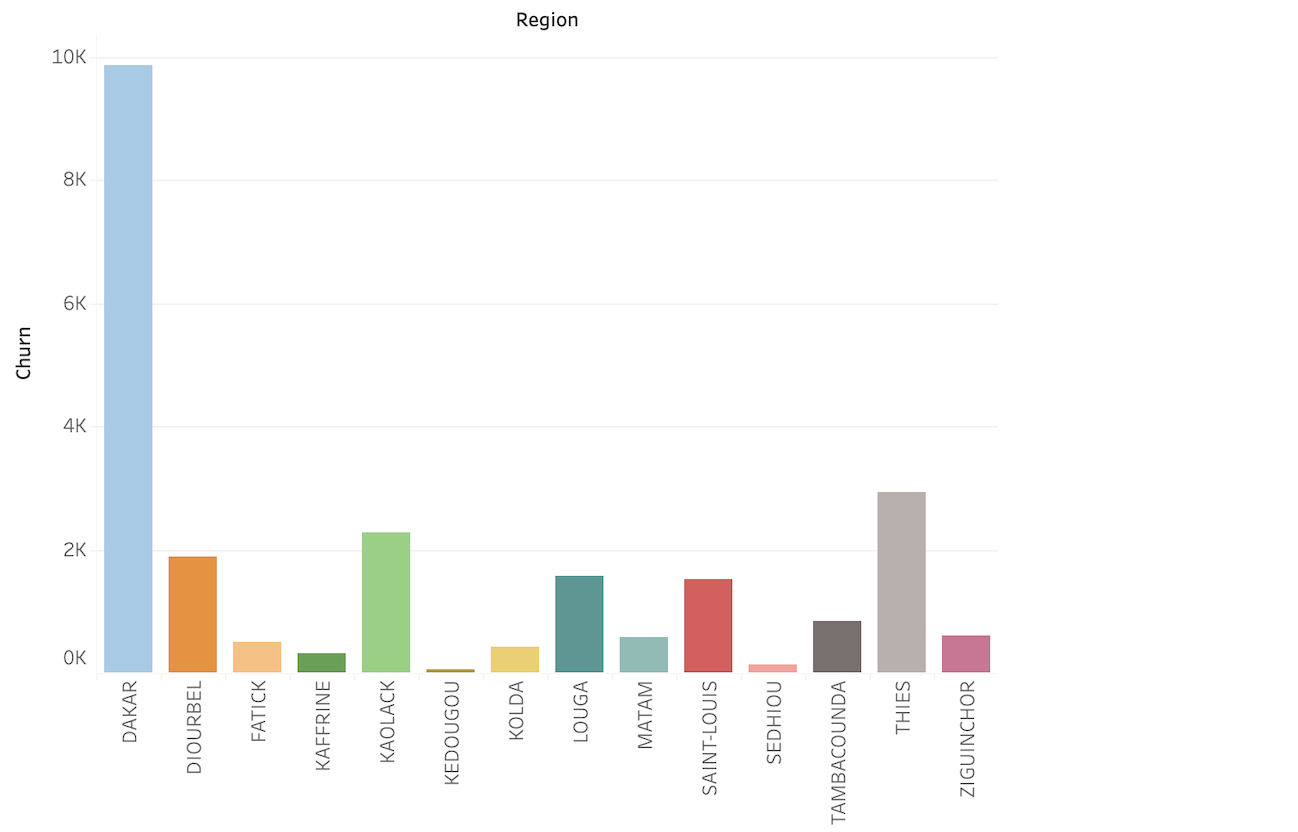
As seen in Fig2, zone 1 and zone 2 columns most of the data (94%) are missing, whereas columns tenure, churn, MRG, regularity, and user\_id don't have any missing values. As seen in Fig3, In KEDOUGOU and SEDHIOU, the number of churn customers is more compared to other regions. While In KAFRFRINE the number of churn customers is very less. According to fig4, for the 12-15 month duration, more customers are churned. Hence, Telecom Company needs to offer better plans for those customers who choose this tenure. On the other hand, very few customers have churned for 3-6 months duration in the network.

Background pattern

Description automatically generatedChart, background pattern

Description automatically generated

Ratio of non-churn(green) and churn(red) customers in each region (Fig3) and tenure (Fig4)

Figure 5 shows the total number of customers for the region. According to the data, the highest number of customers are located in DAKAR and very few customers are located in KEDOUGOU. Figure 6 shows the frequency of customers to churn. So, according to the graph, 80% of the customers are non-churned, while 20% of the customers are churned.Chart, bar chart

Description automatically generated

*Figure 5: Count of customers in each region Figure 6: Frequency of customers to churn*

Figure 7 contains the graphical representation for each column of the dataset, which helps to get the overall idea of the information of each column. Also, we get to know the range and count for each

column from this graphical presentation.

Graphical user interface, application, Word

Description automatically generated *Figure 7: Count of customers in every value range for all features*

## **Attribute Selection**

Few variables are removed and the reason to do so is explained as follows:

* USER\_ID - Unique id for each user value will not affect the output of the model
* ZONE1 & ZONE2 - Around all of the data in these columns are only null values

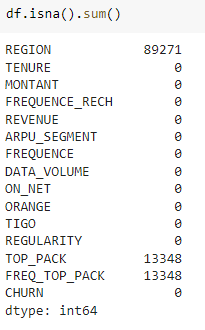


Figure 8: Count of missing values

* MRG - Entire column contains only one value and thus it will have the same weightage to the model throughout

## **Handling Missing Values**

Rows with missing values in columns like MONTANT (Amount), Revenue, Data\_Volume, TIGO (Call count to TIGO), ON\_NET (Inter-expresso call), ORANGE (Call count to ORANGE) are removed. As per the figure 8, there is still a three-column region, top\_pack and freq\_top\_pack have some null values. These three columns will be imputed with constant values in null values. After imputing the constant values, null values from all columns are handled. Columns imputed and values are mentioned in Table 2.

|  |  |
| --- | --- |
| **Variable Name** | **Imputed value** |
| REGION | ‘Others’ |
| TOP\_PACK | ‘No Top Pack’ |
| FREQ\_TOP\_PACK | ‘0’ |

## **Correlation Matrix**

Correlation matrix is a table to show the correlation coefficients between variables. It is plotted between all the variables with numerical values only. The correlation between REVENUE and MONTANT is too high - 0.98 and it is not advised to have two highly correlated variables as it might create a multicollinearity problem. So, the REVENUE column is deleted from the dataset.

Table 2: Imputed values

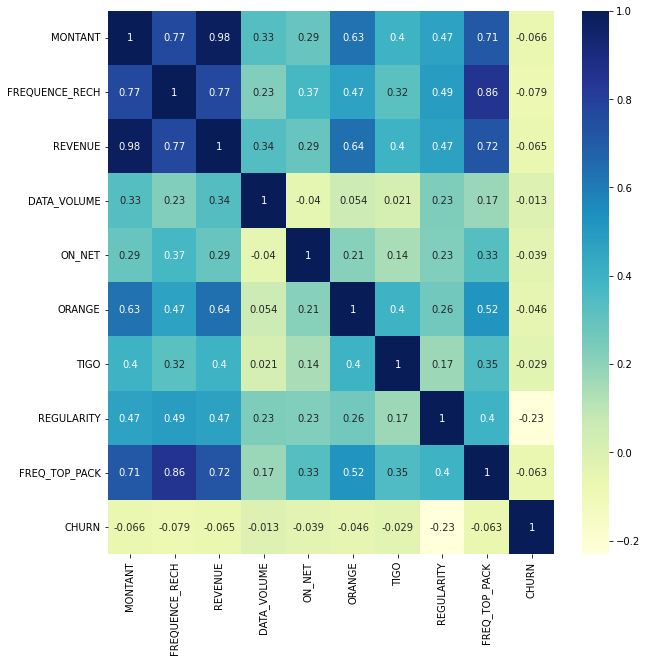


Figure 9: Correlation Matrix

## **Encoding String Variables**

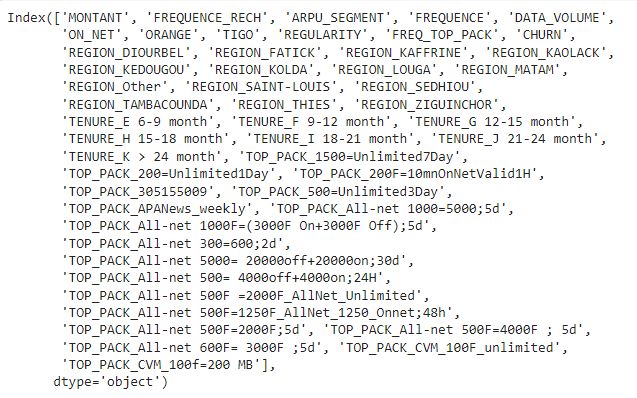
Variables with string values - Region, Top\_Pack, Freq\_Top\_Pack are encoded using OneHotEncoder. Now, the dataset has 137 independent variables

Figure 10: Column names after encoding

## **Data Splitting**

Total records after data cleaning contains 536,226 rows. The cleaned data is split into 80% and 20% of the total records into train and test dataset respectively. Training set consists of 428,980 rows which is used to train the model. Test set consists of 107,245 rows which is used to evaluate the trained model.

## **Feature Scaling**

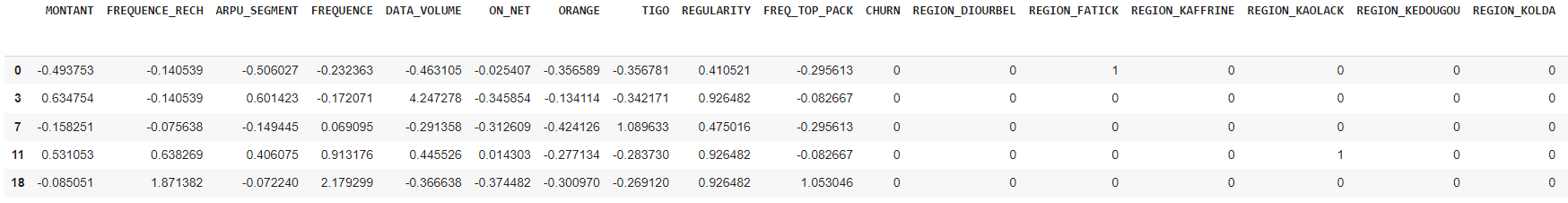
Feature scaling is a method used to normalize the range of independent variables. Since, the values are in different ranges, performing feature scaling will convert the data to a value between -3 and 3 of standard deviation. Feature scaling is applied to both test and train independent variables.

Figure 11: Sample data after feature scaling

## **Over-sampling using SMOTE**

Text

Description automatically generatedIn the training data, the number of records of churn and non-churn customers are in an imbalanced proportion. Using SMOTE algorithm (Synthetic Minority Oversampling Technique), will up-sample the churn customer records to create a balanced sampling. Before over-sampling is performed, the ratio churn to non-churn customers was 40:1 and after over-sampling the ratio becomes 1:1. Only the train dataset is up-sampled and the test dataset is untouched.

Figure12: Output of SMOTE

## **Vector Assembler**

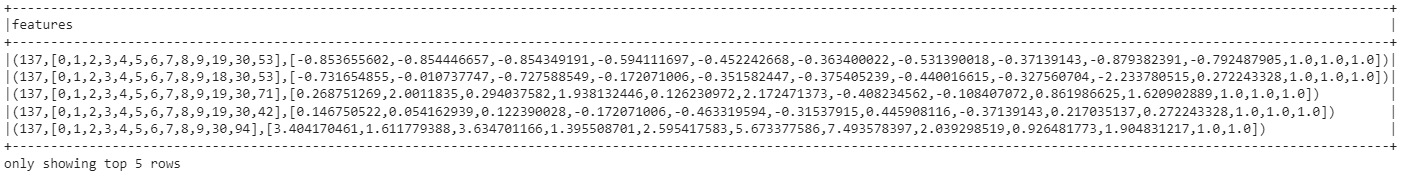
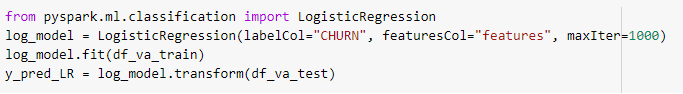
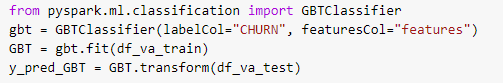
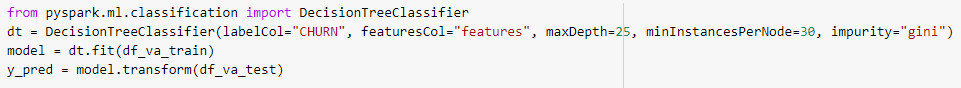
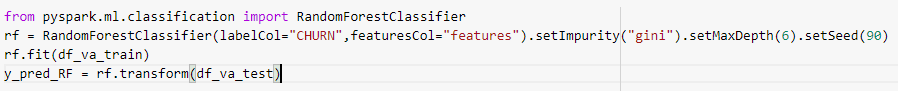
VectorAssembler is a transformer that combines a given list of columns into a single vector column. From the training dataset, the independent variables are added into a single vector column named ‘features’ and are added to the train and test dataset. Dependent variable column is labelled as ‘Churn’.

Figure 13: Features column

# **Modeling**

## **Classification algorithms used**

The target variable is a categorical value i.e. (Yes or No). Therefore, it is the best option to use Classification models to determine the churn. Modelling is performed in Spark using pySpark ML libraries. Classification algorithms used are Decision Tree Classifier, Gradient Boosting Classifier, Logistic Regression classifier, Random Forest Classifier.



## **Principal Component Analysis**

|  |  |
| --- | --- |
| **n\_compenents** | **Cumulative\_Variance** |
| 1 | 58 |
| 10 | 90 |
| 17 | 95 |
| 30 | 98 |
| 96 & above | 100 |

Principal Component Analysis (PCA) is dimension reduction technique for large datasets and also with minimized information loss. Since our dataset contains 137 independent variables, doing PCA might improve model performances. Algorithms are tested with different cumulative variance as mentioned in table 3 and the performance is measured.

Table 3: PCA components and variance

# **Evaluation**

## **Algorithm performance with PCA**

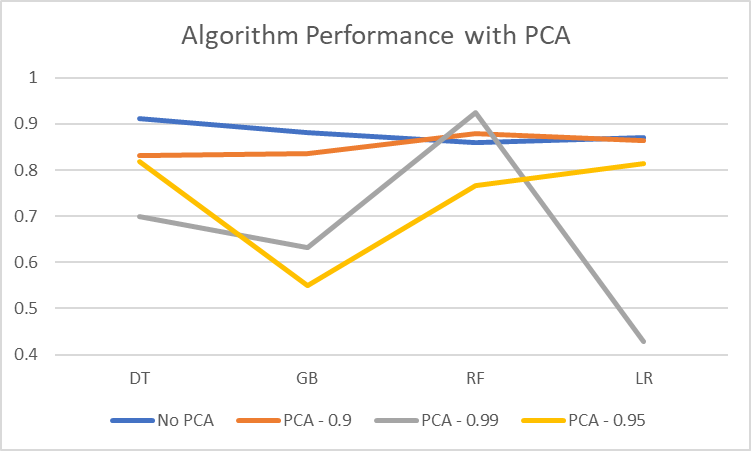
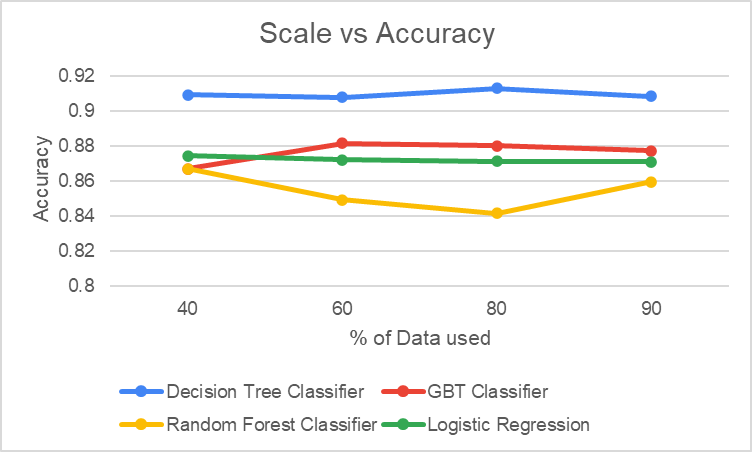
Principal Component Analysis(PCA) has been performed with variance 90, 95 and 99 percentages and the accuracy of each algorithm is measured. All algorithms except Random Forest works better without dimension reduction. Random Forest gave better accuracy with PCA (99% variance). Reasons for not getting better accuracy might be that PCA removed important information and direction of maximum variance is different from class separation direction. So, it is decided to proceed without PCA and the entire cleaned dataset.

Figure 14: Performance of PCA



## **Impact of scale on quality of analysis**

For examining the impact of scale on the quality of analysis, accuracy is selected as the evaluation criteria as the imbalanced effect is already taken into consideration using SMOTE. From the figure 15, the Decision Tree Classifier performs the best with 91% accuracy. The training dataset is split as 40%, 60%, 80%, and 90% for comparing accuracy using the four algorithms. It is observed that for the Random Forest Classifier though the accuracy dips a little for 60 % and 80% of the data size, it almost remains the same for all the three algorithms.

Figure 15

## **Impact of scale on time**

From the figure 16, the execution time goes up with the scale from 40% to 90% of the data size. Decision Tree Classifier performs the best for 90% of the data size with the execution time of 150 sec, whereas Gradient Boost Classifier takes the longest to execute for all the data sizes, especially for 90% of the data size, it takes 380 secs. This might be due to the fact that our dataset has a lot of features.

Figure 16

## Chart, line chart Description automatically generated**Impact of Parallel Computation**

Here AWS EMR clusters with 2 instances (1 Master 1 Slave), 3 instances (1 Master 2 slaves), and 4 instances (1 Master 3 Slaves) are used to compare the performance with the local environment (Specs: 8GB RAM, Quad-Core CPU). Also, this analysis is based on 90% of the dataset as the highest accuracy is obtained for all 4 ML algorithms for this data size. As seen in the fig 17, the local environment takes a longer time than the Amazon Web Services EMR parallel environment for all the cluster sizes. After cluster size 3, time for the Spark jobs flattens and it might further remain flattened for cluster size 5 as well. Also, the local time is almost twice as much as the time for a cluster of size 3. So, overall, cluster size 4 indicates the best performance as it takes the least amount of time as compared to the other 2 clusters.

Figure 17

# **Conclusion**

Insights obtained from the exploratory data analysis are that the majority of the customers who churned are from the city of Dakar, have been with Expresso for more than two years, earned less, and spent less on top up and for data volume. They also made fewer calls to either Expresso, Tigo or Orange lines and did not subscribe to the top active packages. Also, customers who churned are not active more than 6 times in the course of 90 days. After modeling, it is found that there is no definite relationship between dataset size and accuracy. Decision Tree Classifier proves to be the most efficient model in terms of accuracy and time. We suppose the Decision Tree Classifier produced the best results due to the implementation of SMOTE for balancing the dataset. Gradient Boost Classifier requires further hyperparameter tuning that might improve its performance. Also, the server outweighed the Local in terms of time in every ML model demonstrating the power of parallel processors.

# **Recommendation**

Based on the insights from the analysis, Expresso should consider conducting customer micro-segmentation for offering personalized incentives to the targeted customers with the highest tendency to leave. Expresso should also be proactive in addressing the customer issues, feedback and put up a customer engagement strategy for holding back any inactivity on the network. Dissatisfied customers also tend to share negative sentiments regarding customer service on social media channels which are a valuable source of strategic insight. This data can also be leveraged by Expresso to flag service issues that may lead to churning in the future.

# **References**

* https://www.kaggle.com/hamzaghanmi/expresso-churn-prediction-challenge
* <https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html>
* <https://docs.databricks.com/>
* <https://docs.aws.amazon.com/emr/index.html>
* <https://docs.databricks.com/applications/machine-learning/train-model/mllib/index.html>
* <https://towardsdatascience.com/getting-started-with-pyspark-on-amazon-emr-c85154b6b921>
* https://builtin.com/data-science/step-step-explanation-principal-component-analysis

# **Appendix**

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

Chart

Description automatically generated with low confidence