# Detection of Customer Churn

Sahil Sharma

December 5, 2022

|  |  |  |
| --- | --- | --- |
| Introduction | **Contents** | 3 |
| Expected Outcomes |  | 3 |
| Dataset |  | 4 |
| CRISP Model Basis |  | 5 |
| Code Breakdown |  | 6 |
| Exploratory Data Analysis |  | 7 |
| Pipelining |  | 9 |
| Model Selection |  | 9 |
| Predication & Evaluation |  | 10 |
| Summary |  | 12 |
| Result |  | 12 |
| Code |  | 12 |
| Conclusions |  | 13 |
| References |  | 14 |

# Introduction

In this project, we attempt to predict the churn or loss of a customer of a telecommunications company. Having clients leave your company is referred to as customer churn. The concept, designed originally to identify subscribers switching to another service provider within the telecommunications industry, has taken on a significant role in other business domains that include insurance companies, banks, ISPs, etc.

There are two main reasons for customer churn:

* The service provider no longer meets the needs of the customer.
* Customers are offered better incentives by competing service providers, such as better plans, discounts, and offers.

Businesses will be able to isolate potential churners if they have the ability to predict their churn. As part of this process, we will take into account many factors, including the customer's demographics such as age, location, and perceptions about the company. In addition, we will consider browsing patterns, historical data, usage patterns, and support statistics.

# Expected Outcomes

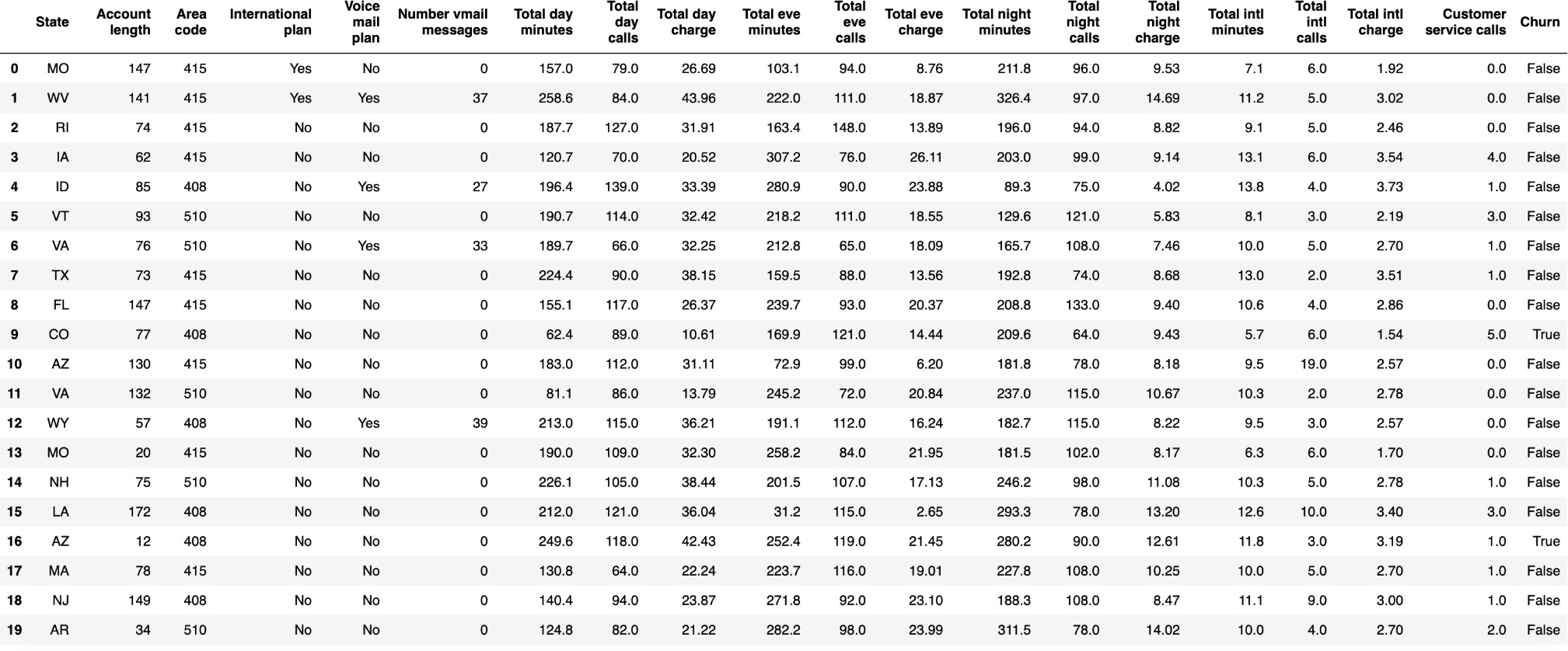
In order for a business to function effectively, it is essential to predict churn. In addition to identifying each client likely to discontinue their service or plan, customer churn prediction would allow the business to personalize its marketing efforts. By offering special offers or plans to identify customers, the business can mitigate this risk. Customers will be less likely to defect, and the costs associated with replacing them will be lower. Every online business has the goal of reducing the churn ratio, which can be achieved through actionable decisions. Two datasets are used in this model to determine whether customers will churn or not, based on input data.

# Dataset

The churn data set from Orange Telecom will be used in this project, namely files churn-80 and churn-20. In the dataset, we will use customer activity data as features, and one column will indicate whether a customer churns or not. Models will be trained with churn-80, while final tests will be conducted with churn-20. Kaggle provided the following data:

# Table 1

*Orange Telecom’s Churn Dataset Schema and sample values:*

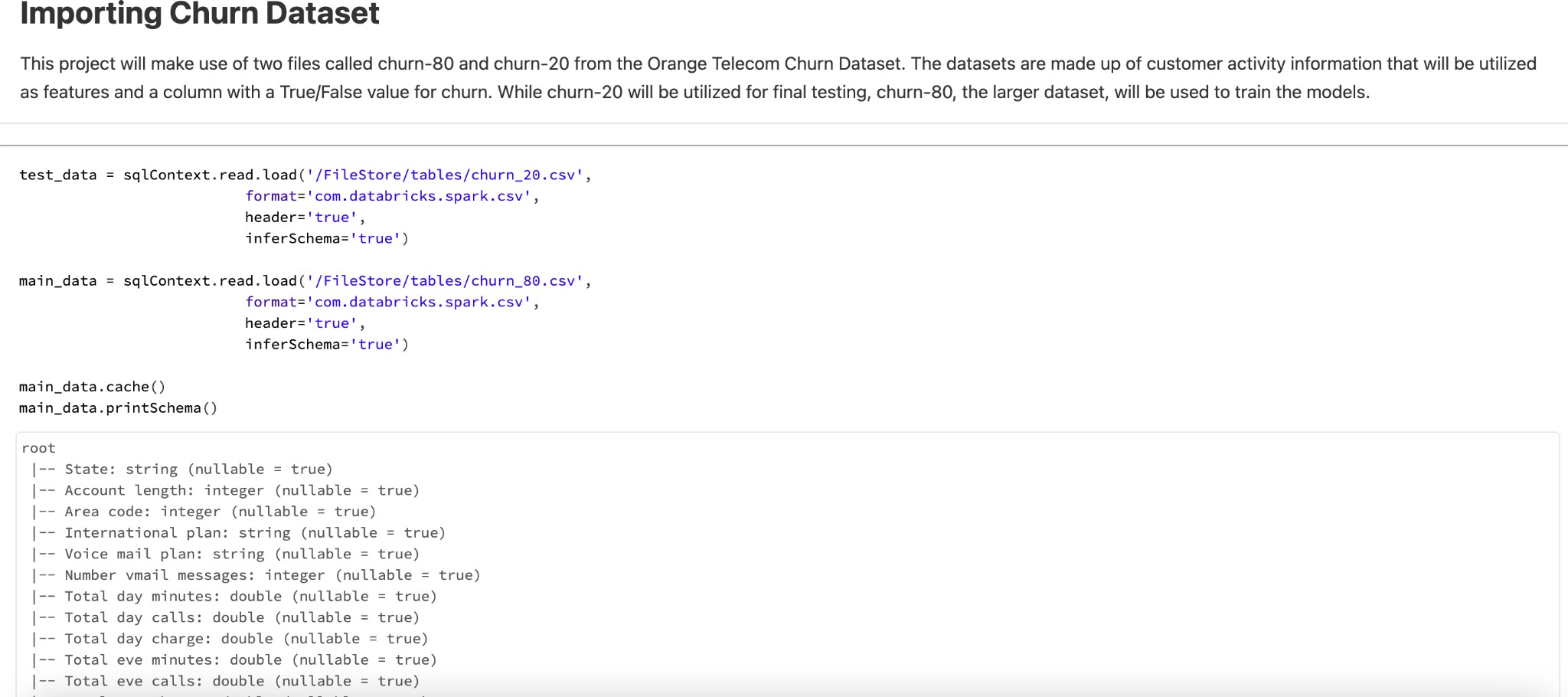


# CRISP Model Basis

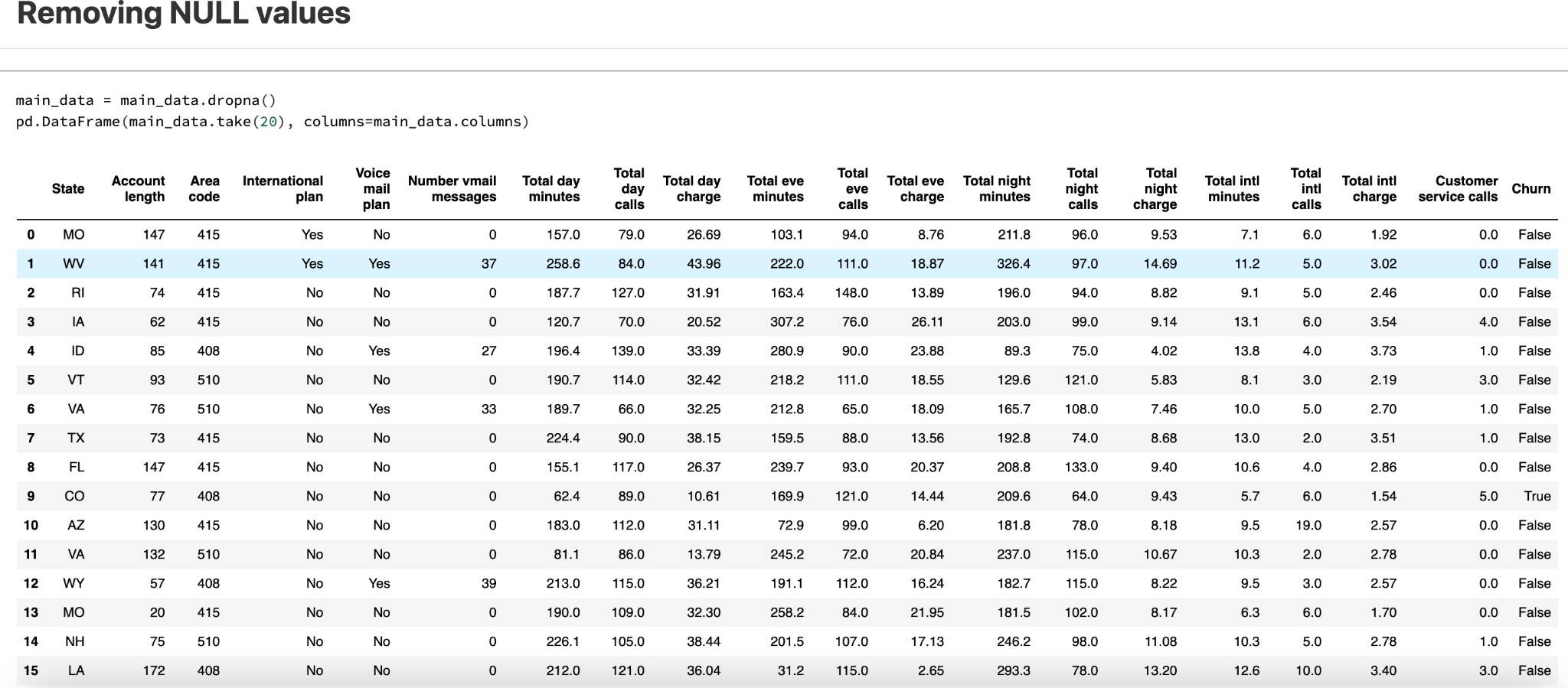
* Business understanding: A telecommunications company, in this case, needs to understand user churn and predict it. It is helpful to know which clients are most likely to stop using the service in the future.
* Data understanding: Our project uses Telecom's Churn Dataset. There are two files in this package: churn-20 and churn-80. The Churn-80 model will be trained and cross-validated with this data. A final evaluation of model performance will be carried out using Churn-20.
* Data Preparation: Any null or missing values were removed from the dataset. Analyzing exploratory data revealed that several features were relevant to the final result (Churn True/False). As a result, columns that were irrelevant were removed and features were transformed.
* Modeling: To evaluate churn prediction, a Decision Tree Classifier was used. In order to determine the most appropriate decision tree model, a pipeline and parameter grid were created.
* Evaluation: Prediction accuracy and F1 score were both considered.
* Deployment: : Databricks Community Edition, with one available cluster, was used to deploy the model. Additionally, the model can be deployed in a variety of ways. Further, AWS or Azure, using REST APIs, is another method that has been proposed

# Code Breakdown

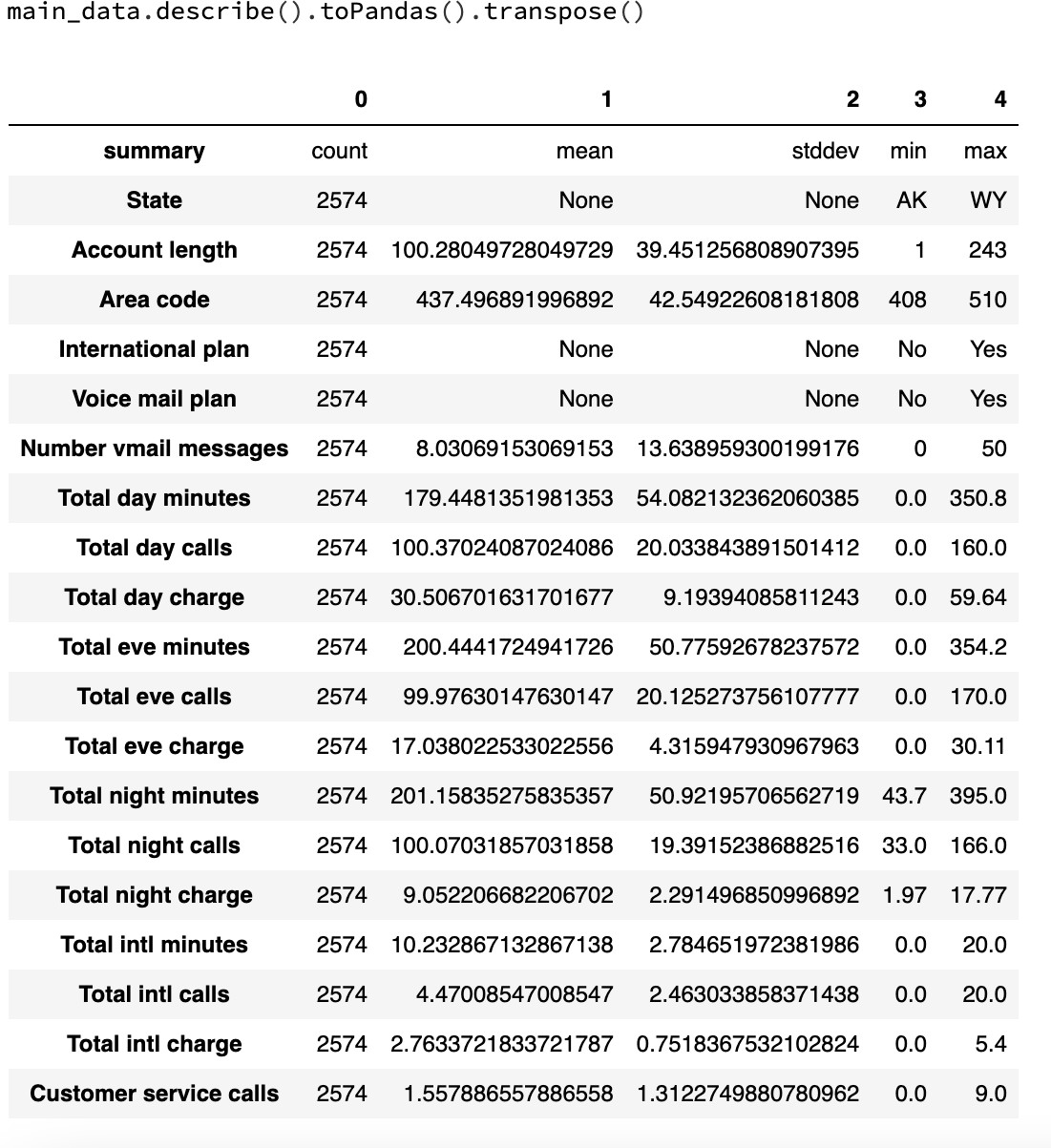
**Figure 1:** *A DataFrame is created by running SQL queries on the data read in with SQLContext.*



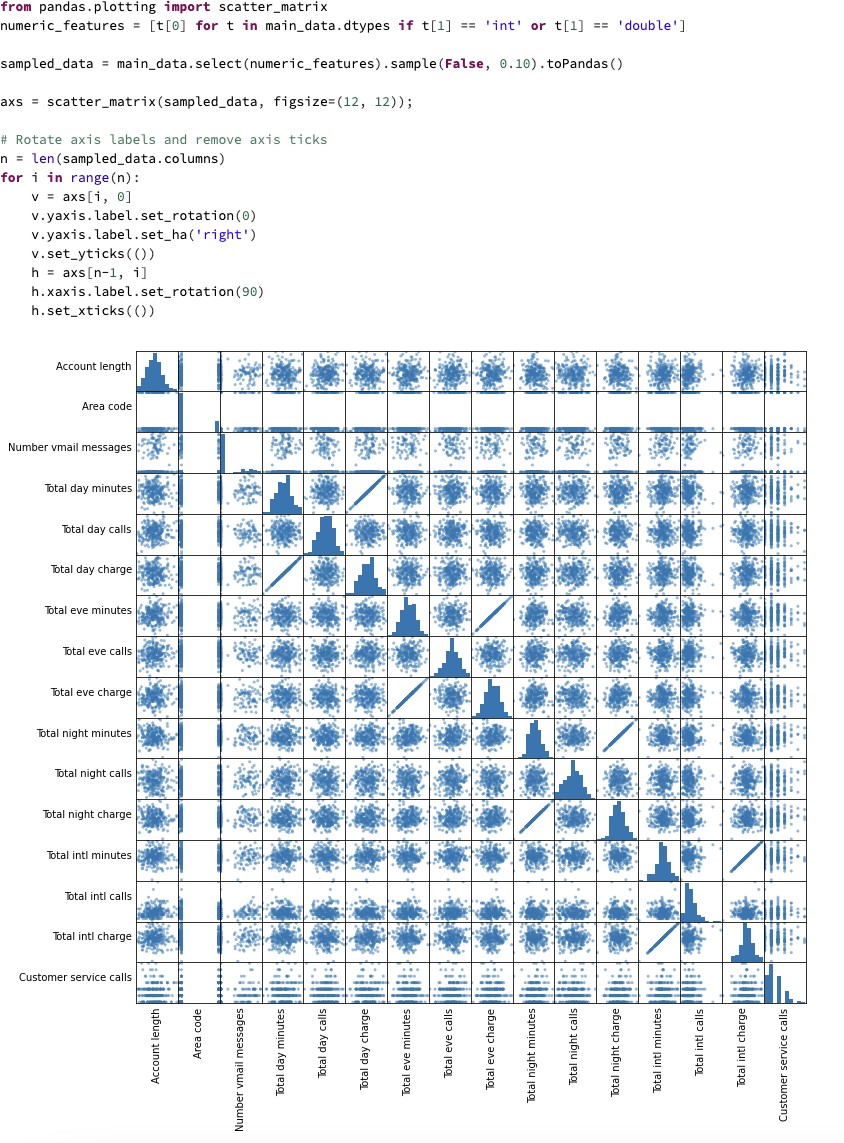
**Figure 2:** *We remove null values from the training dataset when we find them.*



**Exploratory Data Analysis Figure 4:** *We print and observe summary statistics*



**Figure 5:** *Scatter plots are used to observe correlations between columns. An observation is made on a 10% sample instead of using the entire dataset.*



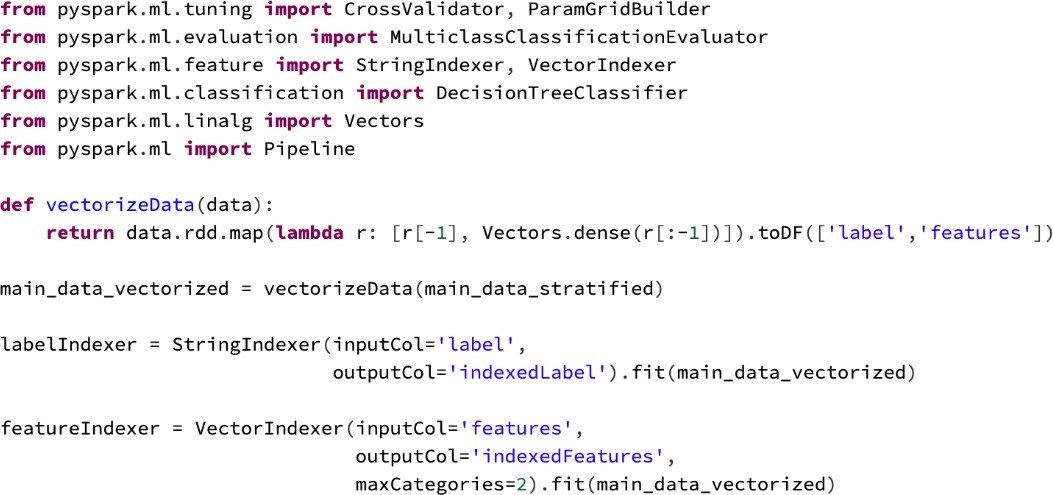
# Pipelining

* Before the model can read the feature columns, they must be vectorized. In this case, we use VectorIndexer().
* A decision tree is being used, so labels and columns of the decision tree are indexed using the StringIndexer() function. Decision trees are therefore better able to handle categorical values because of this.
* The final step is to call the decision tree function.

# Model Selection

* To classify the data in this project, a decision tree classifier is used.
* We create a pipeline that contains DecisionTreeClassifier VectorIndexer, and StringIndexer.
* Various depths of the decision tree are built using a parameter grid.
* Overfitting is mitigated through cross validation. We chose five folds.
* In this case, the F-1 score has been chosen as the evaluation parameter.
* In the end, the best tree model is selected based on the fit of the model to the data.

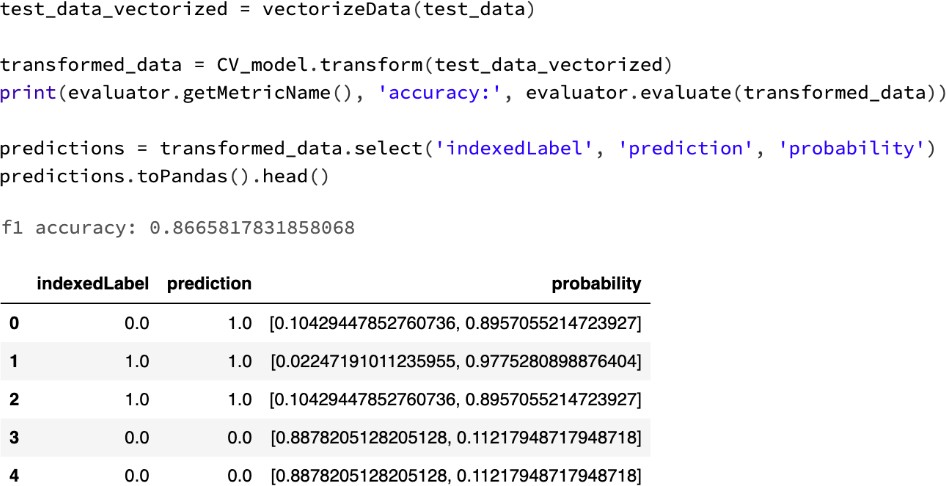
**Figure 7:** *As shown in this figure, the model training code is included.*



# Predictions and Model Evaluation

* The model is evaluated on the testing dataset (churn-20) after it has been fitted to the training dataset (churn-80).
* Model transformations are performed on the vectorized test data after the test data has been vectorized.
* This report prints the prediction results along with the probability associated with each prediction.

**Figure 9:** *This figure illustrates the code and output for model evaluation*



# Summary

A churn prediction model was built in this project. In the training and testing sets, churn-80 and churn-20 datasets were used respectively. To improve our understanding of the data, we cleaned and analyzed it exploratorily. Preprocessed and transformed data were used to create model-readable formats. Prediction and evaluation were performed using a decision tree classifier model. It was determined which decision tree classifier would work best for the dataset using a parameter grid.

# Result

86.6% of the f1 scores were accurate. The resultant predictions of churn and their probabilities were observed.

# Code

[Databricks Published Notebook Link](https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3955085767704038/515851940841170/5009455664670363/latest.html)

# Conclusions

The process of building the project provided various insights into how churn is predicted using Spark. Businesses can use the results and model to determine which customers are likely to churn and make appropriate business decisions based on those results. There is still room for improvement in the project such as:

* There were only a few datasets used. The accuracy of the model can be increased by using larger datasets.
* The primary model used was a Decision Tree Classifier. Besides Decision Trees, there are several other classification models available.
* Databricks Community Edition was used to deploy the model. REST APIs can be used to deploy the model on the cloud, such as AWS or Azure.

# References

* + B. (2016, February 2). GitHub - ben sadeghi/pyspark-churn-prediction: Churn Prediction

with PySpark using MLlib and ML Packages. GitHub: [C hurn Prediction with PySpark](https://github.com/bensadeghi/pyspark-churn-prediction)

[using MLlib and ML Packages](https://github.com/bensadeghi/pyspark-churn-prediction)

* *McDonald, C. (2017, August 4). Churn Prediction With Apache Spark Machine Learning.*

*Dzone.Com.* [*Churn Prediction With Apache Spark Machine Learning - DZone*](https://dzone.com/articles/churn-prediction-with-apache-spark-machine-learnin)

* Machine Learning Model Deployment | DataRobot AI Wiki. (2021, October 22).

DataRobot AI Cloud. Link: [Machine Learning Model Deployment - DataRobot AI Cloud](https://www.datarobot.com/wiki/machine-learning-model-deployment/)

[Wiki](https://www.datarobot.com/wiki/machine-learning-model-deployment/)

* N. (2022, February 7). Machine Learning with PySpark Tutorial. Intellipaat Blog.

[Machine Learning with Spark Tutorial - PySpark MLLIB Tutorial - Intellipaat](https://intellipaat.com/blog/tutorial/spark-tutorial/machine-learning-with-pyspark-tutorial/)

* *PySpark RDD Tutorial | Learn with Examples*. (2022, January 19) [PySpark RDD](https://sparkbyexamples.com/pyspark-rdd/)

[Tutorial | Learn with Examples](https://sparkbyexamples.com/pyspark-rdd/)

* *CRISP-DM*. (2022, April 18). Data Science Process Alliance. Link: [What is CRISP DM?](https://www.datascience-pm.com/crisp-dm-2/)

[- Data Science Process Alliance](https://www.datascience-pm.com/crisp-dm-2/)