**Customer Churn Model**

**Introduction:**

Churn rate (sometimes called attrition rate), in its broadest sense, is a measure of the number of individuals moving out of a collective group over a specific period. The term is used in many contexts, but is most widely applied in business with respect to a contractual customer base. For instance, it is an important factor for any business with a subscriber-based service model, including mobile telephone networks and pay TV operators. Churn rate is an important input into customer lifetime value modeling, and can be part of a simulator used to measure return on marketing investment using marketing mix modeling.

**Data:**

You are provided a dataset that represents a 10,000 customer sample from the European Bank. The columns of this dataset are as follows:

* **CustomerId and Surname:** ID for customer identification and Customer Name respectively
* **CreditScore:** Equivalent to FICO score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Std Dev | Minimum | Maximum |
| 650.53 | 652.00 | 850.00 | 96.65 | 350.00 | 850.00 |

* **Geography:** Data provided for 3 countires: France, Germany and Spain.

|  |  |  |
| --- | --- | --- |
| Row Labels | Count of Observations | Exited (Average) |
| France | 5,014 | 0.162 |
| Germany | 2,509 | 0.324 |
| Spain | 2,477 | 0.167 |
| Total | **10,000** | **0.204** |

* **Gender:** Male and Female

|  |  |  |
| --- | --- | --- |
| Row Labels | Count of Observations | Exited (Average) |
| Female | 4,543 | 0.251 |
| Male | 5,457 | 0.165 |
| Grand Total | **10,000** | **0.204** |

* **Age:** Age of the customer

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Std Dev | Minimum | Maximum |
| 38.92 | 37.00 | 37.00 | 10.49 | 18.00 | 92.00 |

* **Tenure:** Years with the bank

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Std Dev | Minimum | Maximum |
| 5.01 | 5.00 | 2.00 | 2.89 | 0.00 | 10.00 |

* **Balance:** Outstanding balance at the time of snapshot

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Std Dev | Minimum | Maximum |
| 76,486 | 97,199 | 0 | 62,397 | 0 | 250,898 |

* **NumOfProducts:** Number of products (of the same bank) with the client

|  |  |  |
| --- | --- | --- |
| Row Labels | Count of Observations | Exited (Average) |
| 1.00 | 5,084 | 0.277 |
| 2.00 | 4,590 | 0.076 |
| 3.00 | 266 | 0.827 |
| 4.00 | 60 | 1.000 |
| Grand Total | **10,000** | **0.204** |

* **HasCrCard and IsActiveMember:** Indicator for Client having a credit card and indicator for active member respectively
* **EstimatedSalary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Std Dev | Minimum | Maximum |
| 100,090 | 100,194 | 24,925 | 57,510 | 12 | 199,992 |

* **Exited:** Dependent variable (1=Churn, 0 = No Churn)

|  |  |  |
| --- | --- | --- |
| Row Labels | Count of Observations | Percentage of Observations |
| 0.00 | 7,963 | 79.63% |
| 1.00 | 2,037 | 20.37% |
| Grand Total | **10,000** | **100.00%** |

**Logistic Regression:**

Logistic regression is a popular method to predict a categorical response. It is a special case of Generalized Linear models that predicts the probability of the outcomes. In spark.ml logistic regression can be used to predict a binary outcome by using binomial logistic regression, or it can be used to predict a multiclass outcome by using multinomial logistic regression. Use the family parameter to select between these two algorithms, or leave it unset and Spark will infer the correct variant.

**Library**

* from pyspark.ml.classification import LogisticRegression

**Model**

* churn\_model = **LogisticRegression**(featuresCol='features',labelCol='Exited')

**Pipeline**

* pipeline = Pipeline(stages = [geography\_indexer, geography\_encoder, gender\_indexer, gender\_encoder, assembler, churn\_model])

**Train the model**

* fit\_modelfit\_mod = pipeline.fit(train\_data)

**Test the model**

* results = fit\_model.transform(test\_data)

**Accuracy**

* 80.86%

**Decision Tree Classifier**

Decision trees are popular methods for the machine learning tasks of classification and regression. Decision trees are widely used since they are easy to interpret, handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions. Tree ensemble algorithms such as random forests and boosting are among the top performers for classification and regression tasks.

The spark.ml implementation supports decision trees for binary and multiclass classification and for regression, using both continuous and categorical features. The implementation partitions data by rows, allowing distributed training with millions or even billions of instances.

**Library**

* from pyspark.ml.classification import DecisionTreeClassifier

**Model**

* churn\_model = **DecisionTreeClassifier**(featuresCol='features',labelCol='Exited', maxDepth = 8)

**Accuracy**

* 84.78%

**Random Forests**

Random forests are ensembles of decision trees. Random forests combine many decision trees in order to reduce the risk of overfitting. The spark.ml implementation supports random forests for binary and multiclass classification and for regression, using both continuous and categorical features.

**Library**

* from pyspark.ml.classification import RandomForestClassifier

**Model**

* churn\_model = **RandomForestClassifier**(featuresCol='features',labelCol='Exited', numTrees = 60)

**Accuracy**

* 85.82%

**Gradient-Boosted Trees (GBTs)**

Gradient-Boosted Trees (GBTs) are ensembles of decision trees. GBTs iteratively train decision trees in order to minimize a loss function. The spark.ml implementation supports GBTs for binary classification and for regression, using both continuous and categorical features.

**Library**

* from pyspark.ml.classification import GBTClassifier

**Model**

* churn\_model = **GBTClassifier**(featuresCol='features',labelCol='Exited', maxDepth=4)

**Accuracy**

* 86.50%