GENERAL SCOPE DOCUMENT

# INDEX

|  |  |  |
| --- | --- | --- |
| SNo | Name | Page No |
| 1 | Business Use case | 3 |
| 2 | Data Inputs Understanding | 3 |
| 3 | Model Build | 5 |
| 4 | Stratio Components Used | 5 |

# Business Use Case

To build an application that will be used by banks, which will filter good and bad client credit scores when receiving a credit request from the user. A prediction will be produced for each request hitting the application in a near real-time fashion. The credit risk can be identified from the predictions made. As an addition the predictions are labelled as Red, Green and Yellow to denote the credit scores.

# Data Inputs Understanding

There are two datasets (internal and external) used for this use case. The same is uploaded under Data folder for reference.

1. **German Credit**

**Description:** There are two internal datasets used namely german\_credit\_list.csv and german\_credit\_data.csv.

german\_credit\_data.csv is used as a training set for building a model. It has columns mentioned below including the target variable called ‘Risk’ column. The relationship of target variable with other columns are statistically analysed , to select valid feature columns. These feature columns are given as an input to the model. This dataset has a mixture of continuous numerical values and categorical column values.

german\_credit\_list.csv is used as a testing set for executing the model build. Unlike other dataset, this list do not have target variable and it is obtained from the model prediction.

**Columns:**

**Input:** Below are the columns in the input dataset. This is classified as Categorical or Continuous values.

Attribute 1: (numerical)

(continuous values)

      Age in years

Attribute 2: (qualitative)

      Gender

Attribute 3: (qualitative)

(Numeric - Categorical)

      Current Job

      0: unemployed/ unskilled - non-resident

      1: unskilled - resident

      2: skilled employee / official

      3: management/ self-employed/

    highly qualified employee/ officer

Attribute 4: (qualitative)

(Categorical)

      Housing

Attribute 5: (qualitative)

(Categorical)

      Saving Account

Attribute 6: (qualitative)

(Categorical)

      Checking Account

Attribute 7: (numerical)

(continuous values)

      Credit amount

Attribute 8: (numerical)

(continuous values)

      Credit duration

Attribute 9: (qualitative)

(Categorical)

      Purpose

**Output:** A Risk column indicating the credit risk as ‘good’ or ‘bad’. It is a Categorical variable.

1. **External Dataset**

**Description:** This is an external dataset called externalList.csv. The external data set is used for applying business rules and classifying the model predicted risk into different class labels like Red, Green and Yellow.

|  |  |
| --- | --- |
| **Label** | **Condition** |
| Red | Risk Probability >= 0.6 and total fraud >= 2 |
| Green | Risk Probability < 0.15 and total fraud = 0 |
| Yellow | Others |

**Columns:**

The columns of this dataset are binarized and has a value of 1 or 0.

**Input:** Below are the columns in the input dataset

Attribute 1: (numerical – 0 or 1)

LegalCase: The applicant has been involved in a legal case involving unpaid credits.

Attribute 2: (numerical – 0 or 1)

FraudSuspicion: The applicant has been registered in any external data source as fraudulent when paying subscription services like Telecommunications, mobiles, etc.

Attribute 3: (numerical – 0 or 1)

PoliceReport: The applicant is registered in at least one Police Report related to fraud.

Attribute 4: (numerical – 0 or 1)

ContactAudit: The applicant may be in a source of fraud audit data.

Attribute 5: (numerical – 0 or 1)

UkvCheck: The applicant has not paid the income tax.

Attribute 6: (numerical – 0 or 1)

AddressFraudCheck: The applicant has used false addresses fraudulently.

# Model Build

Below are the steps followed in the use-case to build a model.

* Data Analysis and selection
  + Data exploring using graphs and finding the proximity of columns to target risk variable. This will help in understanding which columns needs to be excluded for feature selection

* Data Pre-process
  + Cleaning data - Fixing data values NA with null
* Data Transformation
  + Selected column data are used for data transformation
  + Continuous value column are made categorical keeping certain threshold rules.
  + All categorical columns are scaled by binarizing it to 0 or 1 .
  + Target variable is also binarized

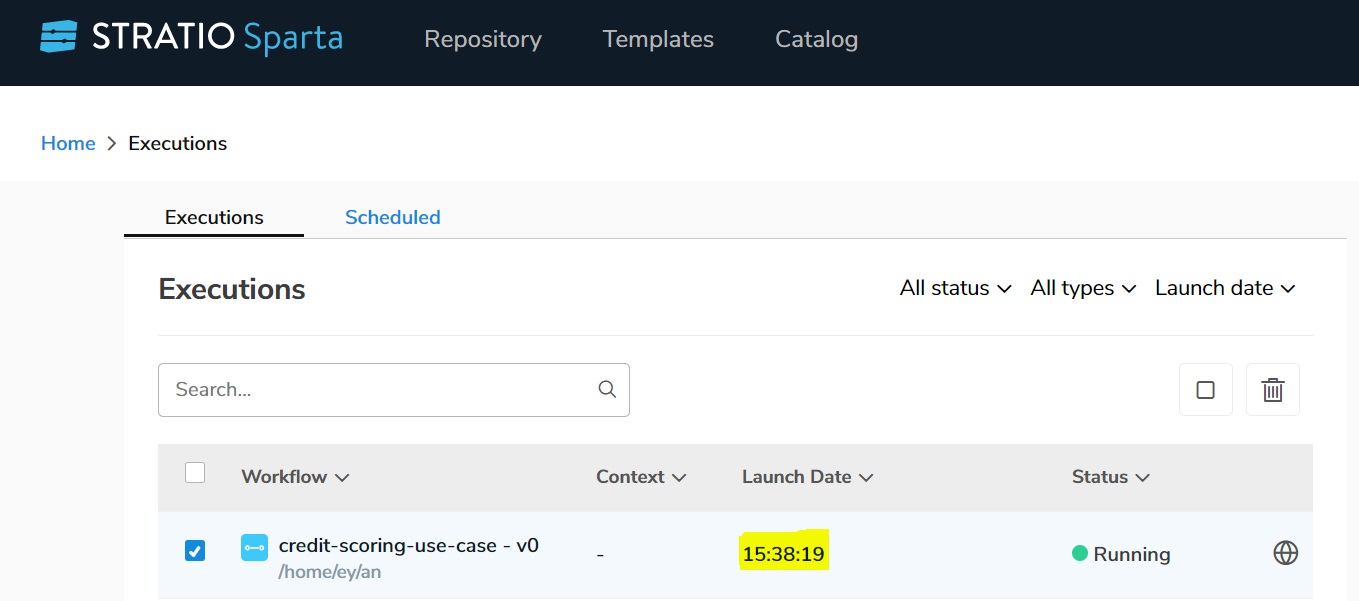
* Model
  + Decision tree machine learning model is implemented.
  + The classification tree analysis capability of this algorithm is utilised where information about data set is gathered and predictions about items target value (here Risk column) is made.
  + The prediction belongs to the class to which the target column belongs.

# Stratio Components Used

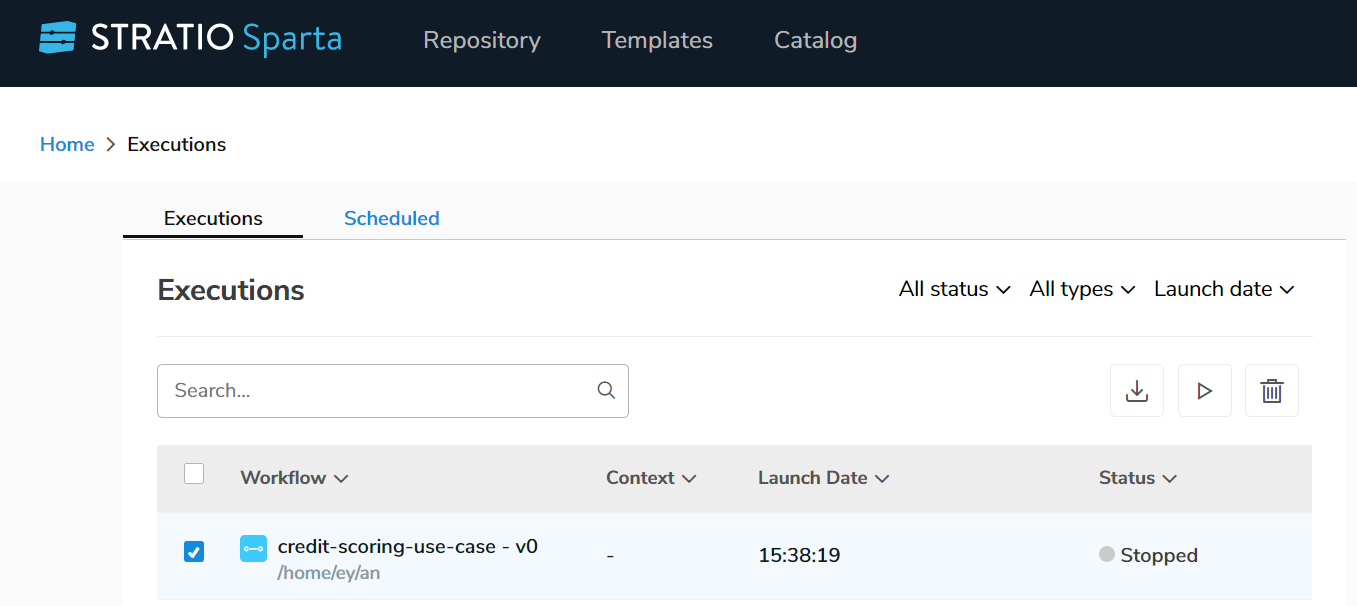
* Stratio Intelligence
  + Python Jupiter notebooks are used for data analysis, pre-processing, transformation and model analysis.
  + PySpark Jupiter notebooks is used for model building and deploying to marathon.
  + The model is saved in the Data Centric platform.
* Sparta Workflow
  + Sparta workflow is used to use this model build from data centric for each of input dataset and labelling them to different credit score classes. The output risk predicted data is saved in HDFS.

Below are certain screenshots when the sparta workflow is deployed using Marathon.

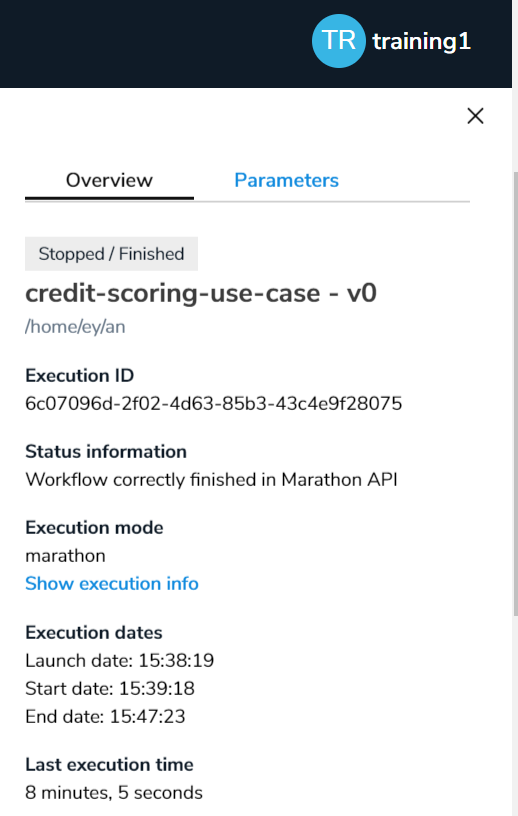
In Running State:



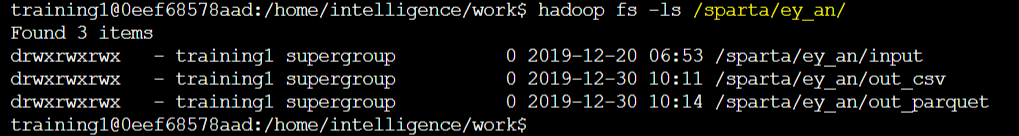
In Completed State



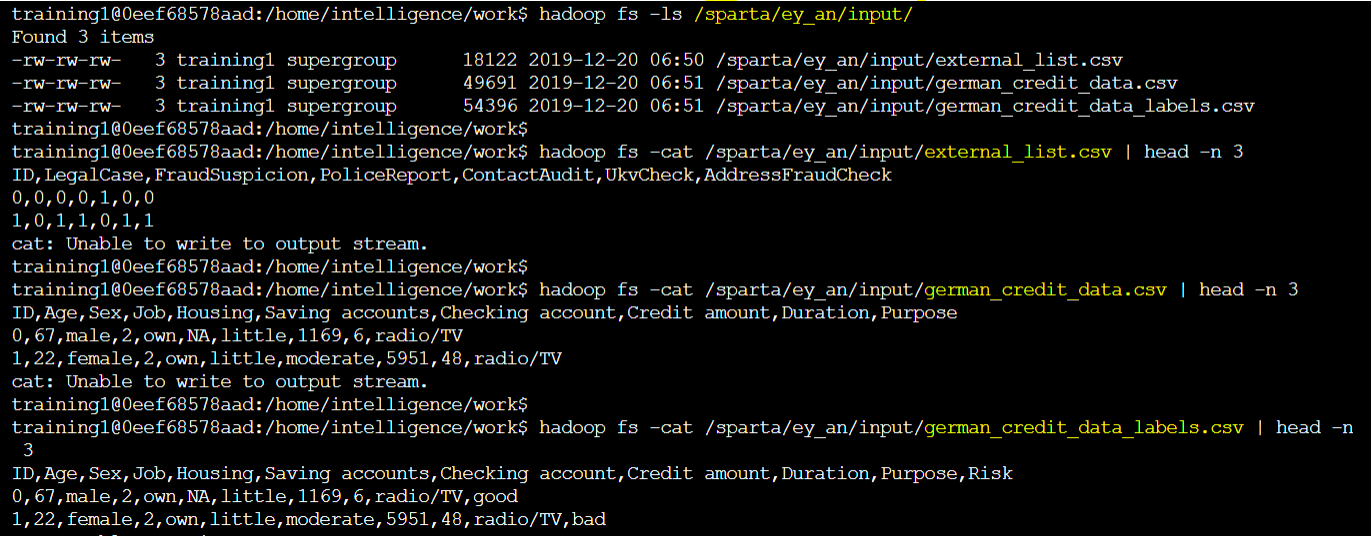
Workflow details



Data in intelligence:



Input



Output

