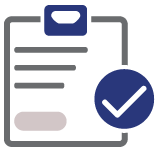
Terminology

The table below outlines the different course activities:

|  |  |
| --- | --- |
| ACTIVITY | DESCRIPTION |
| Demonstration | The Instructor will demonstrate the workflow, outlining the key concept(s). The student is not expected to replicate the Instructors demonstration; but understand the key concept(s) and workflow. |
| Lab | The Instructor will outline the key concepts, features and options. The student is expected to follow along with the instructor so that they understand the key concept(s), features and options for the Exercise |

The icon indicates an Info Tip. Info Tips help users understand unfamiliar workflows or actions.

The icon indicates that you need to be careful when implementing or configuring the step / option(s).

*The icon indicates Best Practice. A Best Practice is a method or technique that has been generally accepted as a standard way of doing things.*

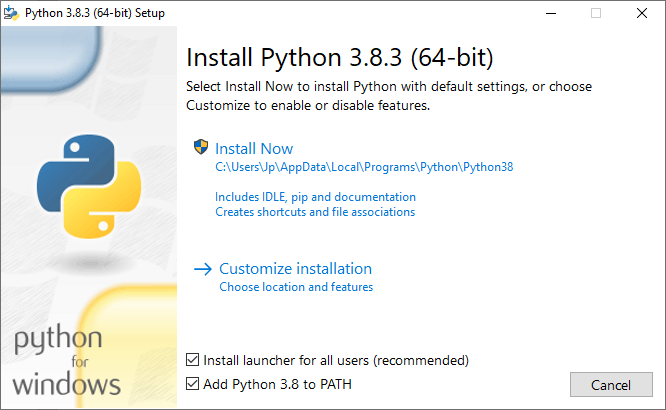
Pre-requisites

The following pre-requisites need to be completed:

* Install Python for Windows
* Google Colab
* Install R for Windows
* Set R Environmental Variables
* Install R Studio for Windows
* Configure Pentaho Data Integration with R

Installation of Python

1. To install: download Python from [Python for Windows](https://www.python.org/downloads/windows/)
2. Click on: the option " Download Windows x86-64 executable installer" on the page.
3. Run: python-3.8.3-amd64.exe file and follow the installation instructions.

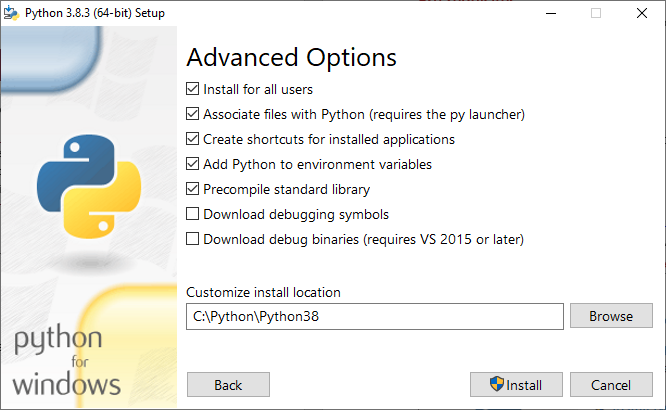


1. Ensure you select:

* Customize Installation
* Install launcher
* Add Python 3.8 to PATH

1. Keep default options and click: Next
2. Ensure you select:

* Install for all users
* Precompile standard library
* Change the Path to: C:\Python\Python38

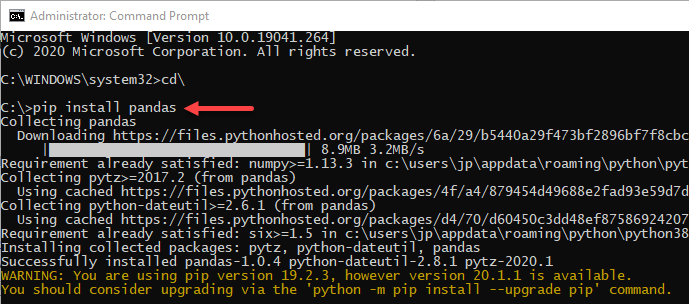


1. Click: Install

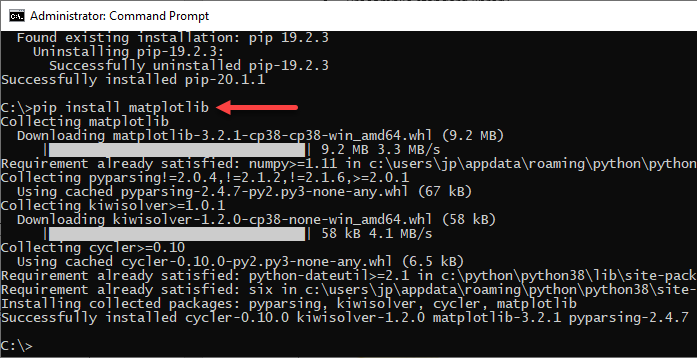
The following libraries need to be installed:

* pandas
* matplotlib
* py4j
* numpy
* wheel
* sklearn
* TPOT

1. Open a Command Prompt (Admin) window
2. Enter the following command to install pandas: pip install pandas



1. Repeat to download and install the other required libraries



You may need to restart your system.

Google Colab

Colab is a Python development environment, based on Jupyter Notebooks, that runs in the browser using Google Cloud. It provides a runtime, fully configured for deep learning libraries, such as, Keras, TensorFlow, PyTorch, and OpenCV.

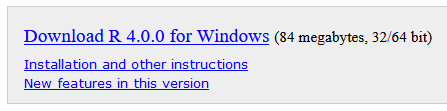
* For further details: [Google Colab](https://colab.research.google.com/notebooks/intro.ipynb)
* Recommended to sync with Google Drive.

Installation of R

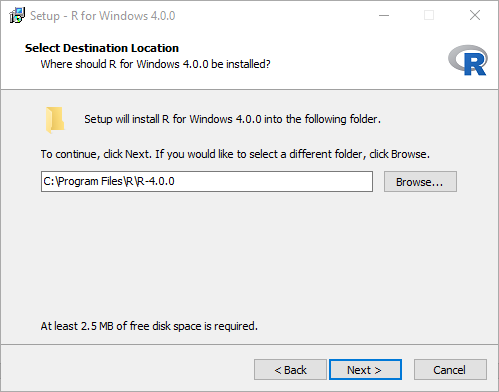
1. To install: download R from [r-project](https://cran.r-project.org/mirrors.html)

* Select a CRAN location (a mirror site) and click the corresponding link

1. Click on: the "Download R for Windows" link at the top of the page.
2. Click on: “install R for the first time" link at the top of the page.

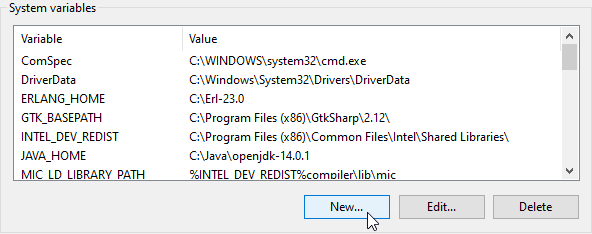


1. Click: Download R <version> for Windows
2. Run: R-4.0.0-win.exe file and follow the installation instructions.

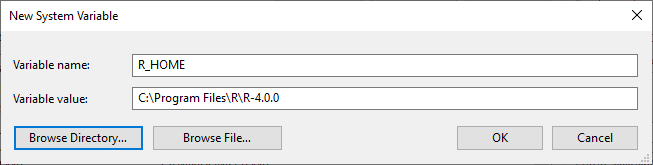


Set R Environmental Variables

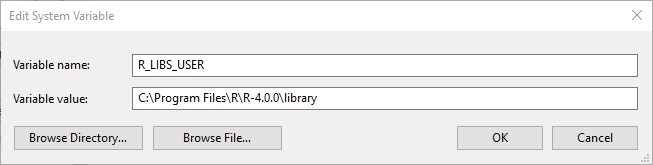
1. Go to: Control Panel > System > Advanced System Settings.
2. Click: Environment Variables button.
3. Under System Variables, click: New.



1. In the Variable Name field, enter: R\_HOME
2. Browse for the directory: C:\Program Files\R\R-4.0.0

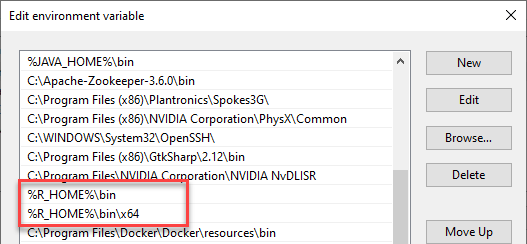


1. Repeat to add the variable: **R\_LIBS\_USER**
2. Browse for the directory: C:\Program Files\R\R-3.6.3\library



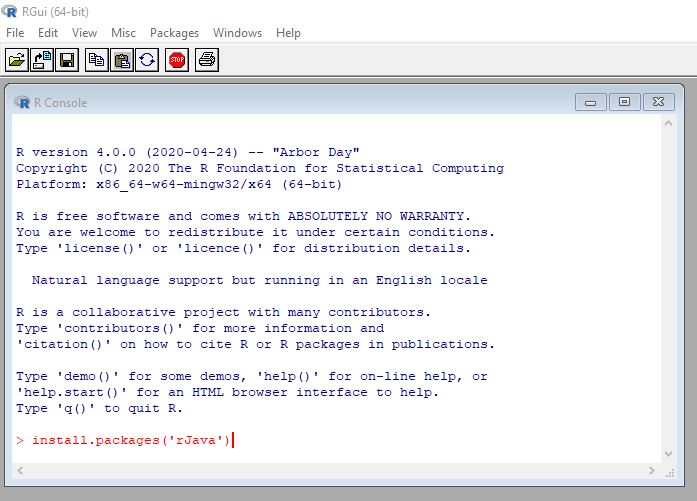
1. Add to the Path the location of the R executable: %R\_HOME%\bin\x64

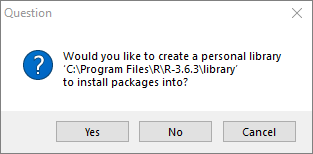
* Ensure the path references rcmd.exe and r.dll



1. Start R. In the R Console: Run the following command:

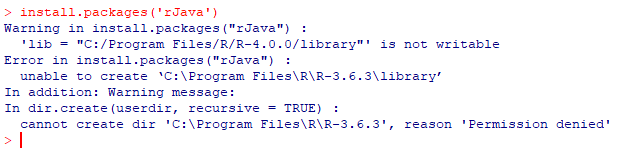
install.packages('rJava')

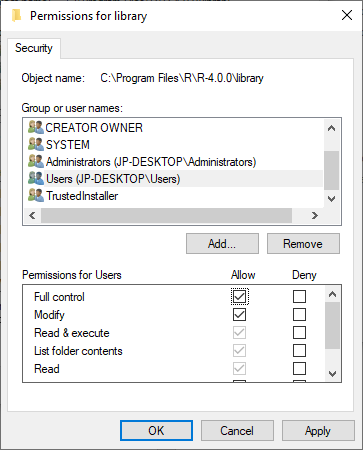


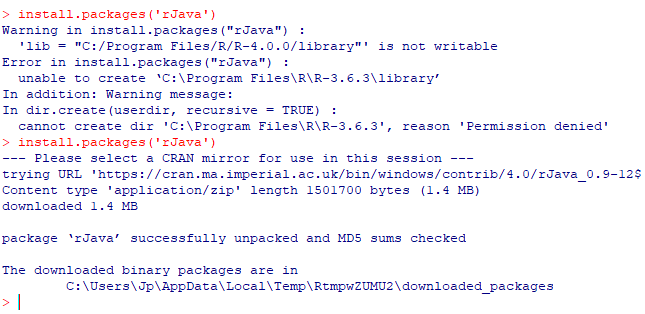
* If prompted with "Would you like to use a personal library instead?" click Yes.
* 
* If prompted with the path of the library, click Yes.
* When prompted for the CRAN mirror, choose a country then click OK.



*You may be denied permission writing to the library folder. You will need to change the permission for the folder.*

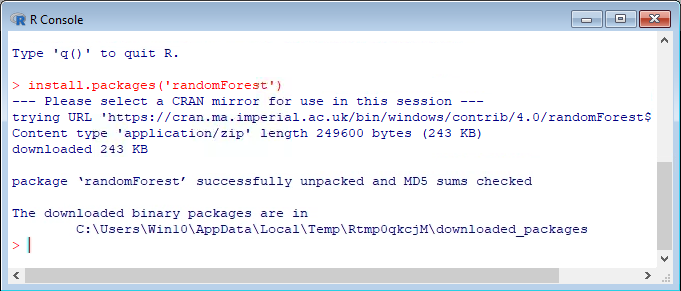






Install randomForest library

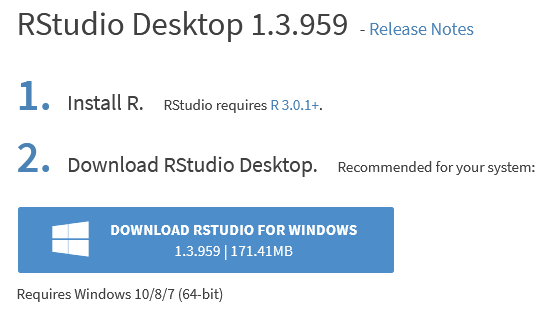
1. Click on: R Console icon
2. Enter the following command: install.packages(‘randomForest’)



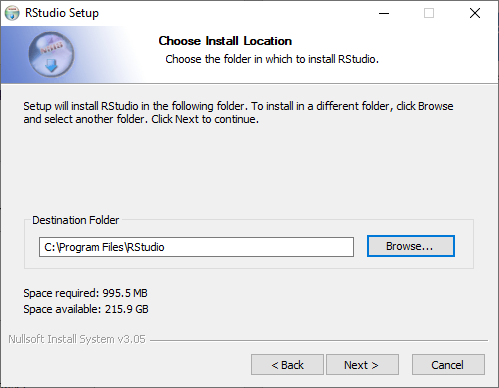
1. After randomForest has successfully been installed, type q() to quit the R console.
2. Click Yes to close the workplace image.
3. Close R.

To Install RStudio

1. To install: download R Studio from [R Studio IDE](https://rstudio.com/products/rstudio/download/#download)
2. Click on the "Download RStudio for Windows" button.



1. Run: RStudio-1.3.959.exe file and follow the installation instructions.



Configure Pentaho Data Integration with R

In the rJava directory there is a jri.dll file that needs to be copied into the libswt directory of Spoon.

1. Stop: Spoon, if it's running
2. Find: %R\_LIBS\_USER%/rJava/jri/x64/jri.dll
3. Copy: jri.dll to the following directory

Windows: ﻿[Pentaho directory]/client-tools/data-integration/libswt/win64

Linux: [Pentaho directory]/client-tools/data-integration/libswt/linux



*Further details can be found at:* [R on PDI](https://support.pentaho.com/hc/en-us/articles/360000307943-Pentaho-Data-Integration#WPR)

Verifying Your Installation

1. Open a new transformation in PDI.
2. Drag an R Script Executor step onto the canvas.
3. Double-click the step and select the middle tab, R Script. You will see some comments at the top of the window:

# The main output is expected to be a data frame, unless "Output

# from script is text" is checked. So, to output a data frame the

# last statement in the script should be the name of the frame.

# In the case that the output is text (as would be seen on the

# R console), the last statement should be a "print" statement in

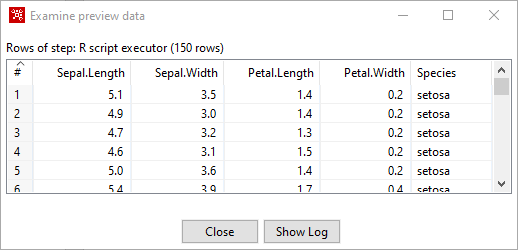
# order to print the object required.

1. Beneath the comment above, enter this code:

library(datasets)

iris

1. Once you have entered this code in the R Script tab, click the **Test Script** button on this tab.



# Lab 1: Credit Card Fraud - AutoML

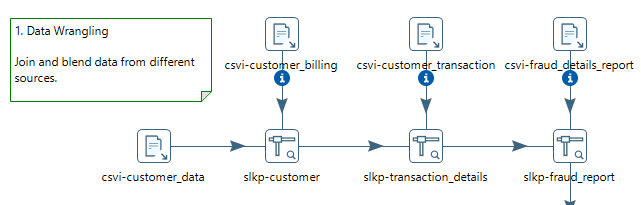
|  |  |  |
| --- | --- | --- |
| Introduction | Imagine that a direct retailer wants to reduce losses due to orders involving fraudulent use of credit cards. They accept orders via phone and their web site, and ship goods directly to the customer. Basic customer details, such as customer name, date of birth, billing address and preferred shipping address, are stored in a relational database.  Orders, as they come in, are stored in a database. There is also a report of historical instances of fraud contained in a CSV spreadsheet. | |
| Objectives | In this guided demonstration, you will:   * Data Prep * Configure Python Executor step. * Build and Train a Forest Tree Model. * Deploy and Test the model. |

Step 1- Data Preparation

With the goal of preparing a dataset for ML, we can use PDI to combine these disparate data sources and engineer some features for learning from it. The following figure shows a transformation demonstrating an example of just that and includes some steps for deriving new fields. To begin with customer data is joined from several data sources, and then blended with transactional data and historical fraud occurrences contained in a CSV file.

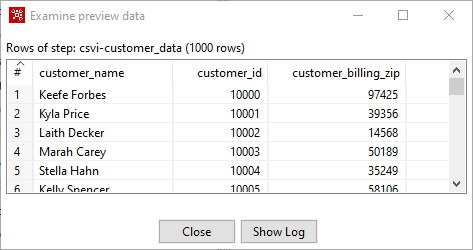
1. In Spoon open the following main Job:

C:\Machine--Learning\Lab\_01\_AutoML\tr\_autoML.ktr



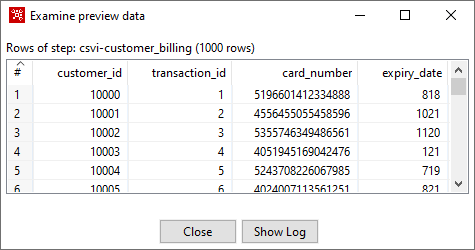
1. Browse the various customer data sources:

csvi-customer\_data



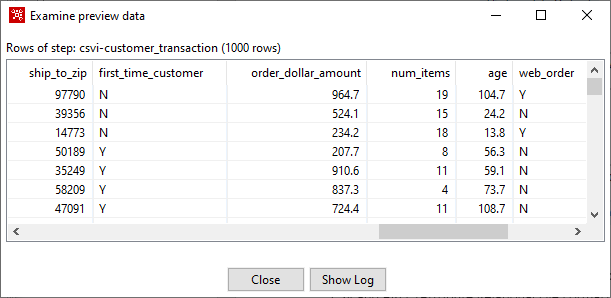
* Here you will find the customer\_billing\_zip codes, which will be used in feature engineering.

csvi-customer\_billing



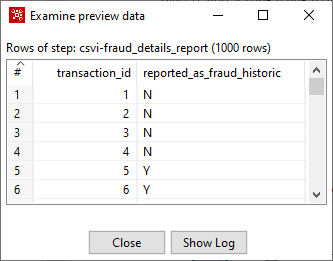
* References the customer transaction

csvi-customer\_transaction



* Customer transaction details
* Feature engineering for ship\_to\_zip
* The transaction details (x variables) are used by the decision trees to determine whether the transaction is fraudulent (y variable). The Boolean values will need to be changed into numbers for the randomForest algorithm.

csvi-fraud\_details\_report

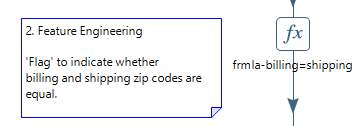


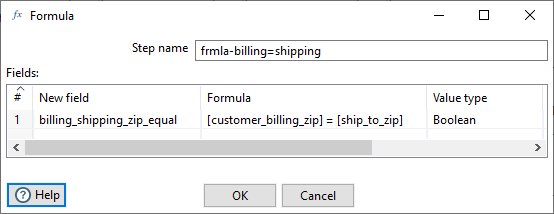
* Indicates whether historically the transaction was fraudulent.

Step 1 - Feature Engineering

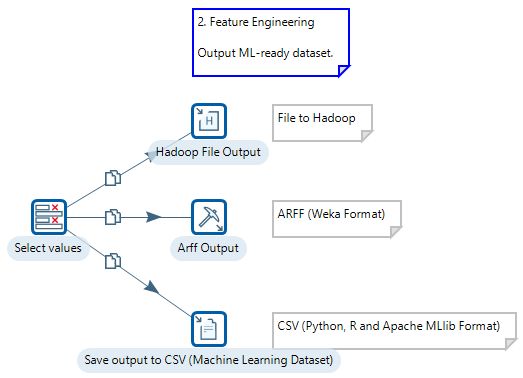
The Feature Engineering is set to: billing zip code = shipping zip code

1. Open the step: frmla-billing=shipping





There are steps for deriving additional fields that might be useful for predictive modeling. These include computing the customer's age, extracting the hour of the day the order was placed, and setting a flag to indicate whether the shipping and billing addresses have the same zip code.



Step 2 - Test Machine Learning Models to Identify the Most Accurate Model

So, what does the data scientist do at this point?

Typically, they will want to get a feel for the data by examining simple summary statistics and visualizations, followed by applying quick techniques for assessing the relationship between individual attributes (fields) and the target of interest which, in this example, is the reported\_as\_fraud\_historic" field.

Following that, if there are attributes that look promising, quick tests with common supervised classification algorithms will be next on the list. This comprises the initial stages of experimental data mining - i.e. the process of determining which predictive techniques are going to give the best result for a given problem.

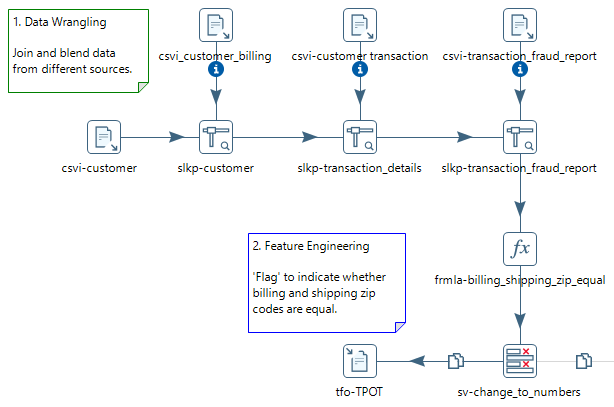
**TPOT: py-auto\_ml**

A Tree-based Pipeline Optimization Tool for Automating Machine Learning (TPOT) is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming. TPOT will automate the most tedious part of machine learning by intelligently exploring thousands of possible pipelines to find the best one for your data.

One of the most useful tools for writing and executing Python script is: [Google Colab](https://colab.research.google.com/notebooks/intro.ipynb)

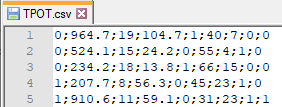
1. Run: transformation:

C:\Machine--Learning\Lab\_01\_AutoML\tr\_autoML.ktr



1. Open the file:

C:\Machine--Learning\Lab\_01\_AutoML\data\TPOT.csv



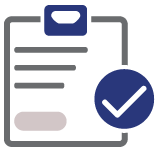
* This will be the dataset used for autoML in Colab.

**Colab**

1. In a Chrome browser, open:
2. <https://colab.research.google.com/notebooks/intro.ipynb>
3. Click on: File > Upload notebook

C:\Machine--Learning\Lab\_01\_AutoML\scripts\credit\_card\_fraud.ipynb

1. Save the Notebook: File > Save

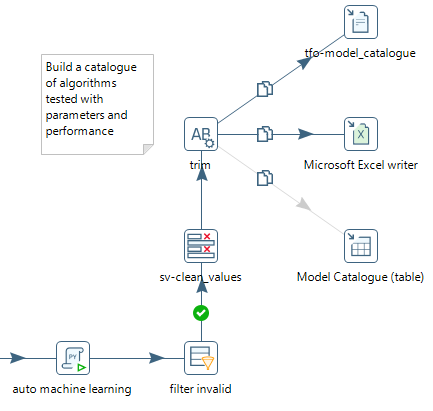
**

*Recommended to Save to either GitHub or Google Drive*

Python Executor

The script has been tested and is now ready for deployment in PDI.

1. Enable the rest of the hops in the transformation, except: Model Catalogue (table)



1. Open the step: py-auto\_ml
2. Ensure to set the path to Python





*For further details on the script see:* **Appendix A**

To ensure the process does not take a long time to process, the following TPOT parameters has been set;

tpot = TPOTClassifier(generations=1, verbosity=2,population\_size=100, config\_dict='TPOT light')



*For further details on TPOT Parameters see:* **Appendix B**

1. Click on the Input tab

* Use this tab to make selections for moving data from PDI fields to Python variables.
* The All rows option is commonly used for data frames. A data frame is used for storing data tables and is composed of a list of vectors of equal length.
* Select the All Rows option to process all your data at once, for example, using the Python list of dictionaries.

|  |  |
| --- | --- |
| Option | Description |
| Available variables | Use the Plus Sign button to add a Python variable to the input mapping for the script used in the transformation. You can remove the Python variable by clicking the X icon. |
| Variable name | Enter the name of the Python variable. The list of Available variables will automatically update. |
| Step | Specify the name of the input step to map from. It can be any step in the parent transformation with an outgoing hop connected to the Python Executor step. |
| Data structure | Specify the data structure from which you want to pull the fields for mapping. You can select one of the following:   * Pandas data frame   The tabular data structure for Python/Pandas.   * NumPy array   The table of values, all the same type, which is indexed by a tuple of positive integers.   * Python List of Dictionaries   Each row in the PDI stream becomes a Python dictionary. All the dictionaries are put into a Python list. |

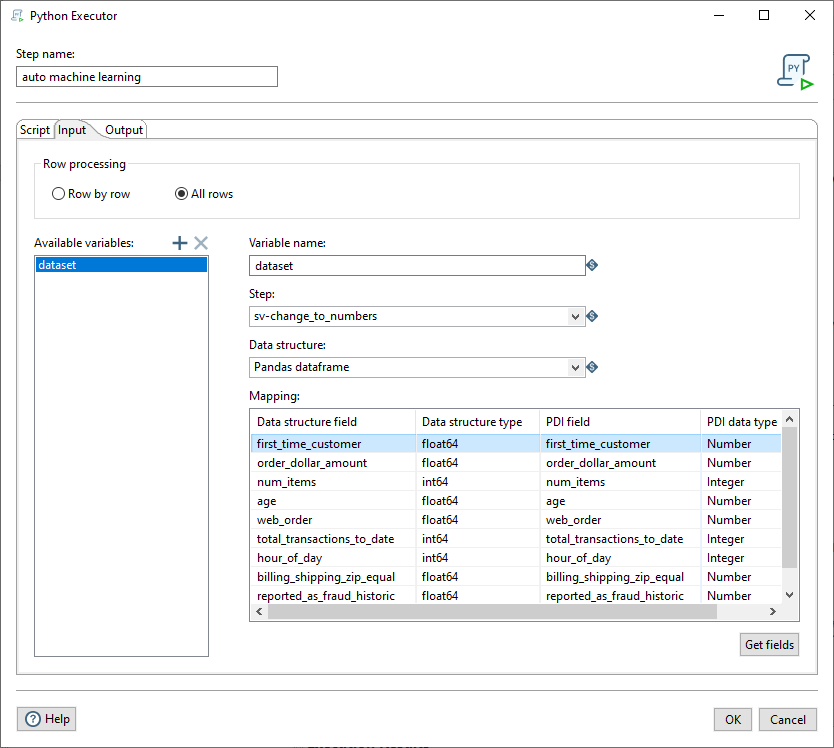
1. The Mapping table contains the following field properties.

|  |  |
| --- | --- |
| Field Property | Description |
| Data structure field | The value of the Python data structure field to which you want to map the PDI field. |
| Data structure type | The value of the data structure type assigned to the data structure field to which you want to map the PDI field. For detailed information on data types. |
| PDI field | The name of the PDI field which contains the vector data stored in the mapped Python variable. |
| PDI data type | The value of the data type assigned to the PDI field, such as a date, a number, or a timestamp. |

1. Select the Get fields button to populate the table with fields from the input step(s) in your transformation. If necessary, you can modify your selections.



*Further details: https://help.pentaho.com/Documentation/9.0/Products/Python\_Executor*



* The cust variable defines the dataframe in the python script using iloc:

x = dataset.iloc[:,1:-1].values

* The dataframe is pulled from the PDI step: sv-changes\_to\_numbers

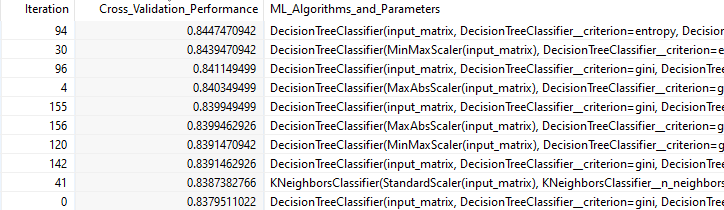
From this list, for the purposes of predictive modelling, we can drop the customer name, ID fields, email addresses, phone numbers and physical addresses. These fields are unlikely to be useful for learning purposes and, in fact, can be detrimental due to the large number of distinct values they contain.

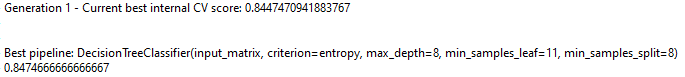
1. Click on: Output tab
2. The output model.df dataframe, from the script:

model\_df=pd.DataFrame(model\_list,columns=['pipe','generation','mutation','crossover','predecessor','operator','cv'])

are converted back to PDI fields

1. Examine the Logging. Sort by Cross validation





The output from the TPOT is ‘tidied’ up before writing the results to a model catalogue file.

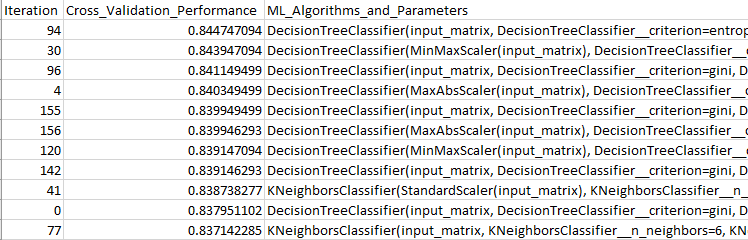
* Filter - removes invalid results
* Select values - orders & renames some of the output fields
* String Operations - trims data stream field
* Text file output - output results to text file: model\_catalogue.txt
* Microsoft Excel Writer - output results to Excel workbook: model\_catalogue.xlsx

What does this mean..?

For the First Generation, the best algorithm pipeline run is: DecisionTree with a scoring of 0.844 and accuracy of 0.84746 (figure used to judge the quality of the pipeline)

1. Open the Excel file:

C:\Machine--Learning\Lab\_01\_AutoML\output\model\_catalogue.xlsx



Conclusion:

The best pipeline to use (with a 85% accuracy) for this dataset is: based on Decision Trees with a min of 8 trees.

May also be worth looking at KNeighbors Classifier

The object of using TPOT is to point you in the right direction for selecting the appropriate algorithm.



*The results will be different each time you run the* TPOTClassifier*.*

# Lab 2: Credit Card Fraud - randomForest

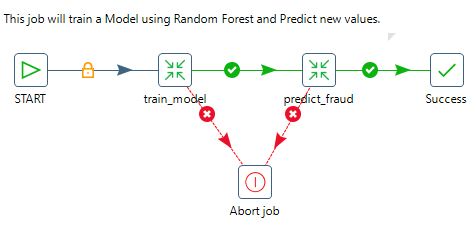
|  |  |  |
| --- | --- | --- |
| Introduction | The results from TPOT point to using a Decision Tree algorithm. | |
| Objectives | In this guided demonstration, you will:   * Train a randomForest model in R * Deploy your model * Predict Fraudulent Credit Card Transactions |

The model that will be used: **randomForest**

**Train the Model**

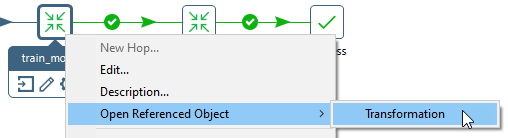
1. In Spoon open the following main Job:

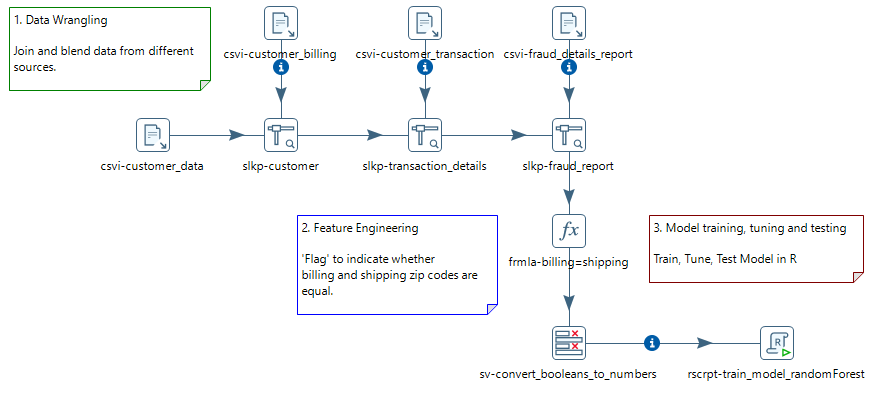
C:\Machine--Learning\Lab\_02\_Credit\_Card\_Fraud\jb\_fraud\_main\_job.kjb



Let’s look at the transformation that Trains for the model.

1. Right mouse click on the train\_model Transformation and select: Open Referenced Object > Transformation





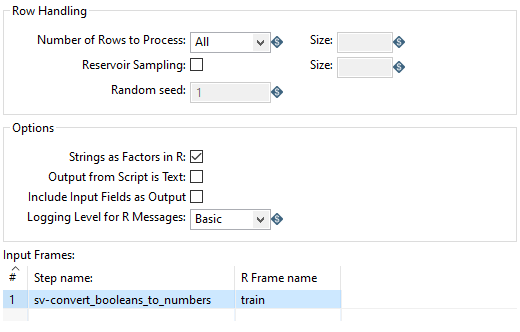
* For an overview of the steps see the previous: Lab1: Credit Card Fraud - AutoML

R script executor: train model

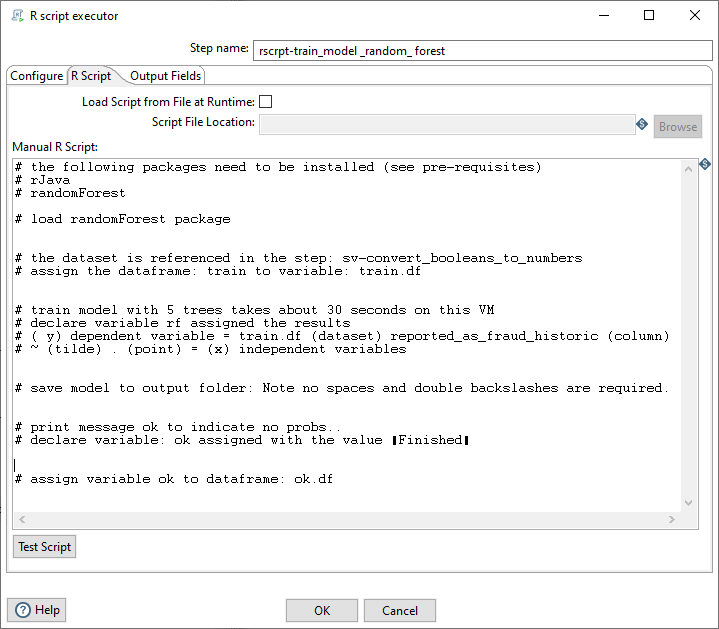
1. Double-click on the ‘rscrpt-train\_model\_randomForest’ step to bring up the configuration settings.
2. Under the Configure tab, ensure the Input Frames points to the Step name:

sv-convert\_booleans\_to\_numbers

the R Frame name: train

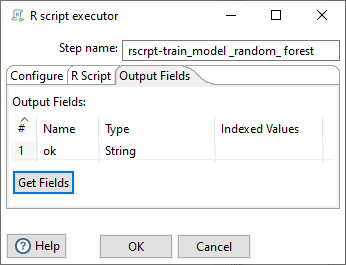


* Processing all records

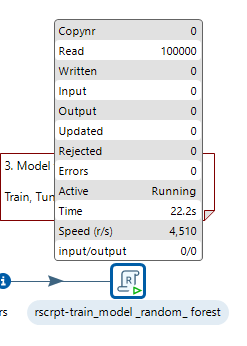
1. Select the R script tab. Copy and Paste the code snippets based on the Comments. 
2. The required script is located:

C:\Machine--Learning\Lab\_02\_Credit\_Card\_Fraud\scripts\train\_model.txt

1. Click on the output tab



1. Run the transformation



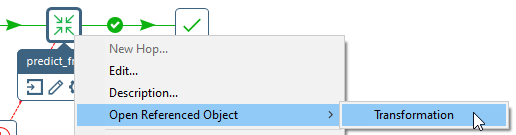
1. Check that the model has been saved in:

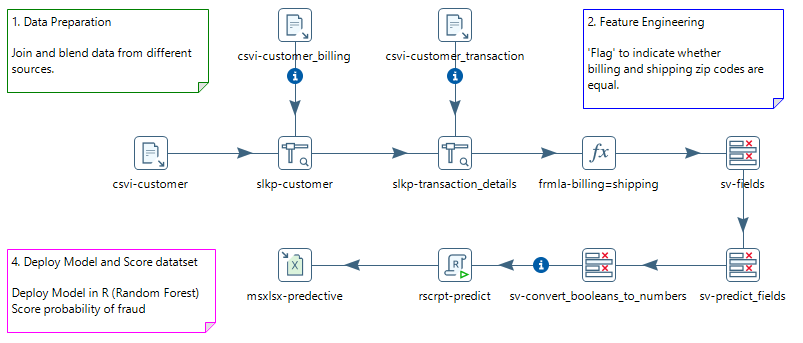
C:\Machine--Learning\Lab\_02\_Credit\_Card\_Fraud\scripts\rf.rdata

**Predict Fraudulent Credit Card Transactions**

Let’s look at the transformation that Predicts fraudulent credit card activity based on our trained model.

1. Right mouse click on the predict\_model Transformation and select: Open Referenced Object > Transformation



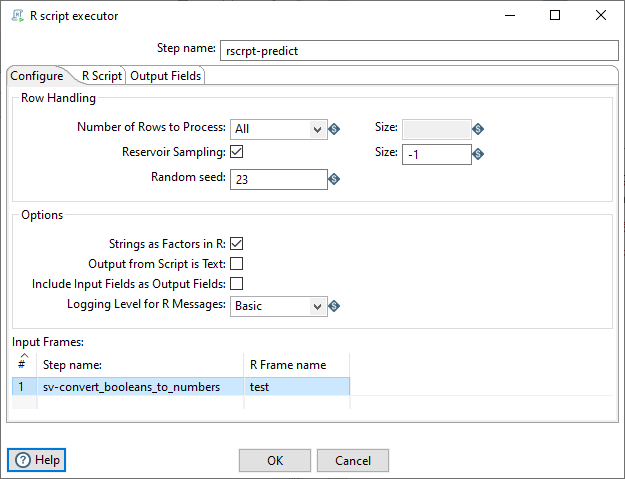


R script executor: Predict Credit Card Fraud

1. Double-click on the ‘rscrpt-predict’ step to bring up the configuration settings.
2. Under the Configure tab, ensure the Input Frames points to the Step name:

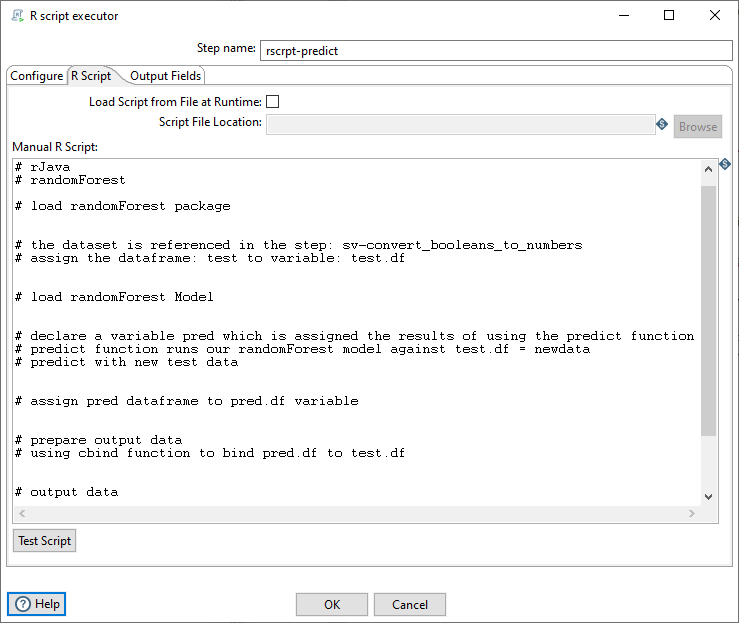
sv-convert\_booleans\_to\_numbers

the R Frame name is: test



* To remove any bias from the dataset, the complete dataset is randomly sampled (mixed up..!)

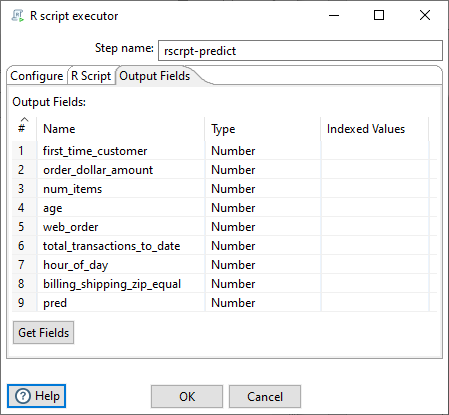
1. Select the R script tab. Copy and Paste the code snippets based on the Comments.



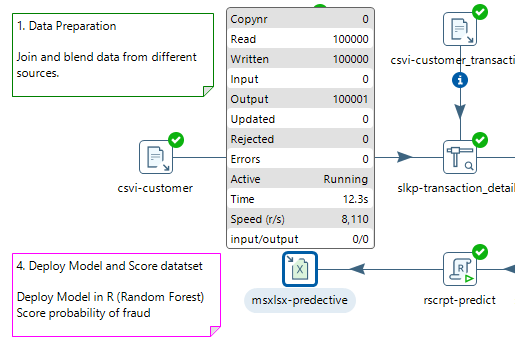
1. The required script is located:

C:\Machine--Learning\Lab\_02\_Credit\_Card\_Fraud\scripts\predict\_model.txt

1. Click on the output tab



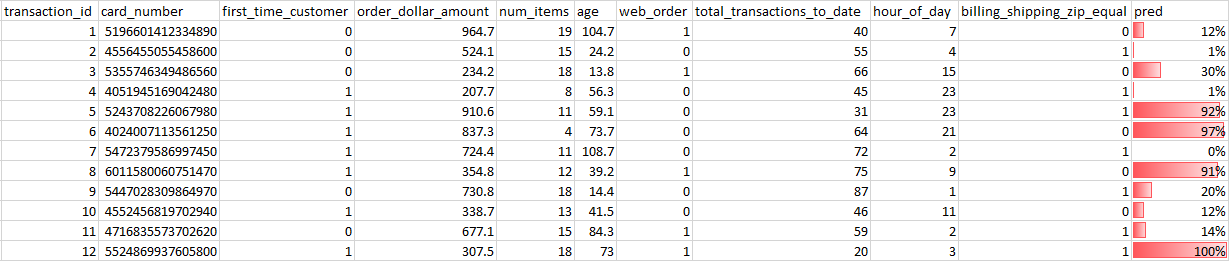
1. Run the transformation



* Ensure all the steps have completed

1. Open the Excel workbook

C:\Machine--Learning\Lab\_02\_Credit\_Card\_Fraud\output\credit\_card\_predict.xlsx



*The complete solution can be found at:*

C:\Streaming--Data\03\_JMS\Lab Apache Artemis\solution

# the following libraries need to be installed (see pre-requisites)

# pandas

# matplotlib

# py4j

# numpy

# TPOT

# import required libraries

import pandas as pd

import numpy as np

from tpot import TPOTClassifier

from sklearn import preprocessing

from sklearn.model\_selection import train\_test\_split

# the dataset is referenced in the step: sv-change\_to\_numbers

# independent variables (x) are referenced starting at row 1: col 1. -1 references all columns apart from the last

x = dataset.iloc[:,1:-1].values

# transform features by scaling each feature to a given range

min\_max\_scaler = preprocessing.MinMaxScaler()

# compute the data minimum and maximum for scaling, then transform.

x\_scaled = min\_max\_scaler.fit\_transform(x)

# optional – change to numpy array

X=np.asarray(x\_scaled)

y=np.asarray(dataset.iloc[:,-1])

# split the dataset into train and test. Test size is set at 75% of dataset (10,000 rows)

# further details on random\_state:

# https://het.as.utexas.edu/HET/Software/Numpy/reference/generated/numpy.random.RandomState.html

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.75, random\_state=None)

# set TPOT parameters (see Appendix: B for further details)

tpot = TPOTClassifier(generations=1, verbosity=2, population\_size=100, scoring='accuracy', config\_dict='TPOT light')

tpot.fit(X\_train, y\_train)

output\_score=str(tpot.score(X\_test, y\_test))

# export TPOT results as python script

tpot.export('tpot\_creditcard\_pipeline.py')

# print the TPOT result

print(tpot.score(X\_test, y\_test))

# PDI output fields which are defined in a dataframe which is mapped to a PDI output field: model\_df

model\_name=[x[0] for x in tpot.evaluated\_individuals\_.items()]

model\_gen=[x[1]['generation'] for x in tpot.evaluated\_individuals\_.items()]

model\_mut=[x[1]['mutation\_count'] for x in tpot.evaluated\_individuals\_.items()]

model\_cross=[x[1]['crossover\_count'] for x in tpot.evaluated\_individuals\_.items()]

model\_predec=[x[1]['predecessor'] for x in tpot.evaluated\_individuals\_.items()]

model\_opp=[x[1]['operator\_count'] for x in tpot.evaluated\_individuals\_.items()]

model\_cv=[str(y[1]['internal\_cv\_score']) for y in tpot.evaluated\_individuals\_.items()]

model\_list=list(zip(model\_name,model\_gen,model\_mut,model\_cross,model\_predec,model\_opp,model\_cv))

model\_df=pd.DataFrame(model\_list,columns=['pipe','generation','mutation','crossover','predecessor','operator','cv'])

Overview of TPOT parameters.

|  |  |  |
| --- | --- | --- |
| Parameter | Valid values | Effect |
| generation | Any positive integer | The number of generations to run pipeline optimization over. Generally, TPOT will work better when you give it more generations (and therefore time) to optimize over. TPOT will evaluate generations x population\_size number of pipelines in total. |
| population\_size | Any positive integer | The number of individuals in the GP population. Generally, TPOT will work better when you give it more individuals (and therefore time) to optimize over. TPOT will evaluate generations x population\_size number of pipelines in total. |
| mutation\_rate | [0.0, 1.0] | The mutation rate for the genetic programming algorithm in the range [0.0, 1.0]. This tells the genetic programming algorithm how many pipelines to apply random changes to every generation. We don't recommend that you tweak this parameter unless you know what you're doing. |
| crossover\_rate | [0.0, 1.0] | The crossover rate for the genetic programming algorithm in the range [0.0, 1.0]. This tells the genetic programming algorithm how many pipelines to "breed" every generation. We don't recommend that you tweak this parameter unless you know what you're doing. |
| num\_cv\_folds | [2, 10] | The number of folds to evaluate each pipeline over in k-fold cross-validation during the TPOT pipeline optimization process. |
| scoring | 'accuracy', 'adjusted\_rand\_score', 'average\_precision', 'f1', 'f1\_macro', 'f1\_micro', 'f1\_samples', 'f1\_weighted', 'log\_loss', 'mean\_absolute\_error', 'mean\_squared\_error', 'median\_absolute\_error', 'precision', 'precision\_macro', 'precision\_micro', 'precision\_samples', 'precision\_weighted', 'r2', 'recall', 'recall\_macro', 'recall\_micro', 'recall\_samples', 'recall\_weighted', 'roc\_auc' or a callable function with signature **scorer(y\_true, y\_pred)** | Function used to evaluate the quality of a given pipeline for the problem. By default, balanced accuracy is used for classification and mean squared error is used for regression. TPOT assumes that any function with "error" or "loss" in the name is meant to be minimized, whereas any other functions will be maximized. |
| max\_time\_mins | Any positive integer | How many minutes TPOT has to optimize the pipeline. This setting will override the generations parameter. |
| random\_state | Any positive integer | The random number generator seed for TPOT. Use this to make sure that TPOT will give you the same results each time you run it against the same data set with that seed. |
| verbosity | {0, 1, 2, 3} | How much information TPOT communicates while it's running. 0 = none, 1 = minimal, 2 = high, 3 = all. A setting of 2 or higher will add a progress bar to calls to fit(). |
| disable\_update\_check | [True, False] | Flag indicating whether the TPOT version checker should be disabled. |

**

*Further details can be found at:* <https://epistasislab.github.io/tpot/using/>

# the following packages need to be installed (see pre-requisites)

# rJava

# randomForest

# load randomForest package

library(randomForest)

# the dataset is referenced in the step: sv-convert\_booleans\_to\_numbers

# assign the dataframe: train to variable: train.df

train.df <- as.data.frame(train)

# train model with 8 trees takes about 50 seconds on this VM

# declare variable rf assigned the results

# ( y) dependent variable = train.df (dataset) reported\_as\_fraud\_historic (column)

# ~ (tilde) . (point) = (x) independent variables

rf <- randomForest(train.df$reported\_as\_fraud\_historic ~ ., train.df, ntree=8, importance=TRUE)

# save model to output folder: Note no spaces and double backslashes are required.

save(rf, file="C:\\Machine—Learning\\Lab\_02\_Credit\_Card\_Fraud\\train\_model\_output\\rf.rdata")

# print message ok to indicate no probs..

# declare variable: ok assigned with the value “Finished”

ok <- "Finished"

# assign variable ok to dataframe: ok.df

ok.df <- as.data.frame(ok)

ok.df

# the following packages need to be installed (see pre-requisites)

# rJava

# randomForest

# load randomForest package

library(randomForest)

# the dataset is referenced in the step: sv-convert\_booleans\_to\_numbers

# assign the dataframe: test to variable: test.df

test.df <- as.data.frame(test)

# load randomForest Model

load("C:\\Machine—Learning\\Lab\_02\_Credit\_Card\_Fraud\\train\_model\_output\\rf.rdata")

# declare a variable pred which is assigned the results of using the predict function

# predict function runs our randomForest model against test.df = newdata

# predict with new test data

pred <- predict(rf, newdata = test.df)

# assign pred dataframe to pred.df variable

pred.df <- as.data.frame(pred)

# prepare output data

# using cbind function to bind pred.df to test.df

submission <- data.frame(cbind(test.df,pred.df))

# output data

submission