

Credit Card Fraud Predictive Modeling and Deployment

**Final Report**

Higher Diploma in Science in Data Analytics

Word Count: 7500

Author: Ciaran Finnegan / 10524150

E-mail: [10524150@mydbs.ie](mailto:10524150@mydbs.ie) / [ciaran@feefinnegan.com](mailto:ciaran@feefinnegan.com)

Supervisor: Dr Shahram Azizi Sazi

Report Submission Date: 25th September 2020

**Abstract**

This is, at maximum, a quarter page summary of the project…

**Acknowledgments**

I wish to acknowledge the..

# Contents

[Contents 2](#_Toc50665554)

[1. Introduction 4](#_Toc50665555)

[1.1. What the Project Aimed to Deliver 4](#_Toc50665556)

[1.2. How the Project Delivery was Implemented 4](#_Toc50665557)

[2. Background / Literature Review 5](#_Toc50665558)

[2.1. Credit Card Fraud Detection: Further Research on Predictive Models 5](#_Toc50665559)

[2.2. Credit Card Fraud Detection: In Context – My Dataset 6](#_Toc50665560)

[3. Requirements: Specification and Design 7](#_Toc50665561)

[3.1. High Level Project Requirements 7](#_Toc50665562)

[3.2. Project Architecture Diagram 7](#_Toc50665563)

[3.3. High Level Project Design 8](#_Toc50665564)

[3.3.1. Prototype Development 8](#_Toc50665569)

[3.3.2. Final Project Deliverable 8](#_Toc50665570)

[4. Project Implementation (1) – Azure Modelling 9](#_Toc50665575)

[4.1. The Machine Learning Workflow 9](#_Toc50665576)

[4.2. Credit Card Fraud – The Azure Workspace/Machine Learning Studio 11](#_Toc50665577)

[4.3. Credit Card Fraud Dataset – Analysis and Preparation 14](#_Toc50665578)

[4.3.1. Experiment 1: Data Cleansing 15](#_Toc50665579)

[4.3.2. Experiment 2: Feature Engineering 16](#_Toc50665580)

[4.3.3. Experiment 3: Feature Selection 17](#_Toc50665581)

[4.4. Credit Card Fraud – Building the Azure Model 18](#_Toc50665582)

[4.4.1. Experiment 4: Basic Model Evaluation with Feature Engineering 19](#_Toc50665584)

[4.4.2. Experiment 5: Model Evaluation with Cross Validation/Hyperparameter Tuning 21](#_Toc50665585)

[4.4.3. Experiment 6: Comparison of Multiple Classification Algorithms (1) 23](#_Toc50665586)

[4.4.4. Experiment 7: Comparison of Multiple Classification Algorithms (2) 25](#_Toc50665587)

[4.5. Credit Card Fraud – Deploying the Azure Model 26](#_Toc50665588)

[4.5.1. Experiment 8: Feature Engineering on Larger Dataset 27](#_Toc50665590)

[4.5.2. Experiment 9: Creation of Predictive Fraud Model for Deployment 28](#_Toc50665591)

[4.5.3. Deployment and Validation of Web Service for Predictive Fraud Model 31](#_Toc50665592)

[5. Project Implementation (2) – Shiny R Dashboard UI 33](#_Toc50665593)

[5.1. Data Visualisations in a Shiny Dashboard 33](#_Toc50665594)

[5.2. Credit Card Fraud – UI to Check Fraud Predictions 33](#_Toc50665595)

[5.3. Shiny UI – Hosted Application 33](#_Toc50665596)

[6. Testing and Results 34](#_Toc50665597)

[6.1. User Story ‘Demos’ – Test Results and ‘Feedback’ 34](#_Toc50665598)

[6.1.1. User Story 4: Initial Data Modelling – Review and Evaluation 34](#_Toc50665602)

[6.1.2. User Story 5: Basic Shiny App – Review and Evaluation 35](#_Toc50665603)

[6.1.3. User Story 6: Integrated Prototype – Review and Evaluation 36](#_Toc50665604)

[6.1.4. User Story 7: Enhanced Modelling – Review and Evaluation 37](#_Toc50665605)

[6.1.5. User Story 8: Enhanced UI – Review and Evaluation 38](#_Toc50665606)

[6.1.6. User Story 9: Presentation Preparation – Review and Evaluation 39](#_Toc50665607)

[6.2. Final Project Assessment 40](#_Toc50665608)

[6.3. Project Plan 2020: Final Status – 25th September 2020 41](#_Toc50665609)

[7. Project Location and User Guide 42](#_Toc50665610)

[7.1. Credit Card Fraud Application: Prototype Location 42](#_Toc50665611)

[7.2. Credit Card Fraud Application: User Guide (Final Project) 42](#_Toc50665612)

[8. Project Conclusions 43](#_Toc50665613)

[8.1. Where Project Goals Achieved? 43](#_Toc50665614)

[8.2. Future Design/Deployment Considerations 43](#_Toc50665615)

[9. Appendices 44](#_Toc50665616)

[9.1. Azure Generated Code Segments 44](#_Toc50665617)

[9.2. Shiny R Application Code Files 46](#_Toc50665618)

[9.2.1. Diagram: The RStudio Cloud Environment 46](#_Toc50665624)

[9.2.2. The Shiny UI Code 46](#_Toc50665625)

[9.2.3. The R Code Parsing Data and Invoking UI 46](#_Toc50665626)

[9.3. Azure Machine Learning Classic Studio Experiments 47](#_Toc50665627)

[9.3.1. Experiment 1: Breakdown 47](#_Toc50665629)

[9.3.2. Experiment 2: Breakdown 48](#_Toc50665630)

[9.3.3. Experiment 3: Breakdown 48](#_Toc50665631)

[9.3.4. Experiment 4: Breakdown 48](#_Toc50665632)

[9.3.5. Experiment 5: Breakdown 48](#_Toc50665633)

[9.3.6. Experiment 6: Breakdown 49](#_Toc50665634)

[9.3.7. Experiment 7: Breakdown 50](#_Toc50665635)

[9.3.8. Experiment 8: Breakdown 51](#_Toc50665636)

[9.3.9. Experiment 9: Breakdown 51](#_Toc50665637)

[9.4. Credit Card Fraud Datasets 52](#_Toc50665638)

[10. References / Bibliography - Interim Report 53](#_Toc50665639)

# Introduction

## What the Project Aimed to Deliver

The artefact at the end of this project is an application that invokes a bespoke predictive model and provides a user with an online interface to retrieve a score for whether a given credit card transaction is likely to be fraudulent.

The Interim Report…

<image>

## How the Project Delivery was Implemented

The Interim Report…

The project .

* The entire remaining 100K+ rows of the credit card fraud dataset will be used to generate the final predictive model.

# Background / Literature Review

## Credit Card Fraud Detection: Further Research on Predictive Models

Section 2.2 of the Interim Report on this project elaborated on two Kaggle submissions made in related to the credit card fraud dataset generated by the work of the Machine Learning Group (<http://mlg.ulb.ac.be>) of ULB (Université Libre de Bruxelles).

References in both those submissions referred to ongoing studies in the area of credit card fraud detection that are being collected by the ResearchGate network for scientists and researchers.

*Figure: www.researchgate.net/project/Fraud-detection-with-machine-learning*

A screenshot of a cell phone

Description automatically generated

The latest submission, as of September 2020, contains an interesting paper of credit card fraud detection with a focus on transaction sequence. However, the initial sections of this submission (Lucas et al., 2019)**n** provide an excellent overview of the challenges facing credit card detection in the real world and solutions that have emerged over the last 10+ years.

Reading through this material I have drawn on key observations to direct my work on this project.

*Algorithm Selection*

A paper from Bhattacharyya, Jha, Tharakunnel and Westland, 2011**n** described research on a real-world US credit card dataset. It involved a comparison of Support Vector Machine, Random Forest, and Logistic Regression, which – as expected - are all algorithm options I have access to in Azure ML Studio (classic).

There are many observations in the paper but some key points I took note of (and are repeated in other articles) were:

* Credit card data is often very imbalanced. Fraud can be disastrous when it happens but it a tiny proportion of overall records. A defined sampling approach is a definite requirement.
* The Fraud/non-Fraud imbalance can make the use of ‘Accuracy’ in a Confusion Matrix as somewhat ineffective.
* Accurate identification of fraud is often a primary requirement so there is a need to look at the trade offs in improving Recall and Precision.
* Logistic Regression can perform consistently well but is dependent on the approach to Feature Engineering.

One interesting note came from other research is that the imbalanced nature of credit card data makes fraud a candidate for anomaly detection routines (Ceronmani Sharmila et al., 2019)**n**. Algorithms such as ‘Isolated Forest’ or ‘Local Outlier Factors (LOF)’ are frequently recommended.

As I explain in Section 2.2 of this document, I choose not to adopt these unsupervised approaches and focused much of my time in the Machine Learning workflow for this project on Feature Engineering leading into supervised learning techniques.

*Feature Engineering*

Although I was not able to read all the details in the paper by Mahmoudi and Duman, 2015**n** on fraud detection analysis, several commentators on this study referred to the benefit of being able to work with the ‘raw’ features of a credit card dataset.

This is an advantage I have with my dataset, as opposed to the previously mentioned ULB dataset, which is heavily anonymised through PCA.

However, the opening lines of a paper from Lima and Pereira, 2017 on ‘*Feature Selection Approaches to Fraud Detection in e-Payment Systems*’ states that “..*Due to the large amount of data generated in electronic transactions, to find the best set of features is an essential task to identify frauds*.“

Given that my starting dataset has 380 columns, this was a guiding principle for me.

I was also going to have to code in R to invoke an API to call my predictive fraud model with all the ‘important’ features on ‘new’ credit card transactions passed as parameters. Therefore, reducing the complexity of setting up this parameter list for the API code would help improve robustness of the UI code.

*Transaction Sequences*

The ResearchNet articles provided references to additional papers on how to improve Feature Engineering for fraud analysis by creating aggregates and time series analyses of the transactions. I choose not to explore this avenue because of the potential complexity.

There are many columns in my dataset that look at time since transaction but my primary actions with this data was just to remove any highly correlated features.

## Credit Card Fraud Detection: In Context – My Dataset

To give an overview of my credit card dataset:

1. It contains 25,128 rows and 380 columns.
2. This is a live dataset of North American credit card transactions from 2013. Only names and initial address lines have been anonymised. Apart from data cleansing, no other alterations to the ‘raw’ transactions have taken place.
3. Previously, the data was used for a, now discontinued, credit card fraud product that relied on a ‘Rule Engine’ to generate alerts for potential fraud.
4. The data is expected to be free of corrupt data elements, and largely free of missing data.
5. Many columns still present in the dataset were created as the result of ETL processes from other peripheral systems and are redundant. No domain knowledge in this area has been documented.
6. Approximately 16.2% of the dataset records are known fraud cases. The data has been balanced over a period of time based on the 2013 dataset. This is a significant advantage/difference from other comparable research datasets in the public domains.
7. The original project proposal, as described in the Interim Report, was to use a dataset of 280K records but that became infeasible due to reasons elaborated on in Section 8 of this document (Conclusions).
8. 10% of the dataset was used for initial feature engineering in Azure ML Studio (classic), but the full dataset was used to train the production model.

Based on the research described in the previous section (2.1), my approach to building this predictive credit card fraud model will focus on:

* Data sampling and balancing and can be relatively straightforward for my project. Other industry papers have devoted significant amounts of time to addressing the challenge of balancing a very small sub-set of actual fraud data.
* Feature Engineering will be very important. This is true of most Machine Learning problems but I need to reduce the number of features from a starting number of 380.
* Algorithm selection, to build my predictive fraud model, can focus on binary Classification options for supervised learning.

# Requirements: Specification and Design

## High Level Project Requirements

To provide an synopsis of the project requirements detailed in the Interim Report..

* A predictive model for Credit Card fraud detection ..
* A ..
* All ..

## Project Architecture Diagram

*Figure: High Level Application Architecture Diagram*

A close up of a map

Description automatically generated

## High Level Project Design

One-page description with a diagram…

The Interim report provided a detailed overview of the User Stories used to map out the design and implementation of this project.



### Prototype Development

*Initial Basic Modelling in Azure ML Studio (Classic)*

The,,

<image>

*Basic UI Deployment*

The..

<image>

### Final Project Deliverable

*Enhanced Modelling in Azure ML Studio (Classic)*

The,,

<image>

*Enhanced UI Deployment*

The..

<image>



# Project Implementation (1) – Azure Modelling

## The Machine Learning Workflow

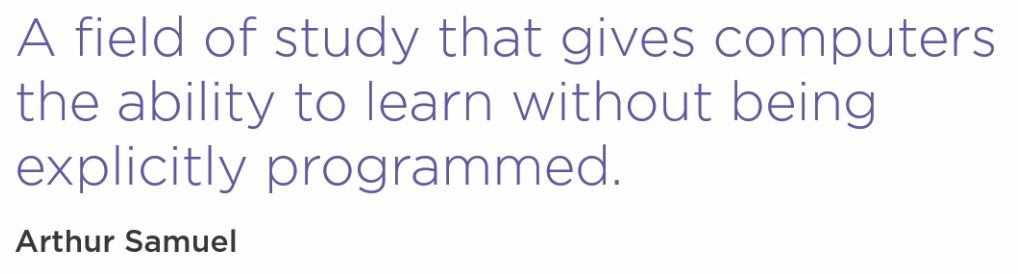
A significant amount of training and reference material I had access for guidance on using the Azure Machine Learning Studio came from Pluralsight courses.

I had reproduced a number of illustrations from those courses (and cited the sources) to;

* Explain my general approach to using Machine Learning processes to build my credit card Fraud predictive model.
* Describe how Azure Machine Learning Studio was used to implement the key steps in the Machine Learning process for this project.

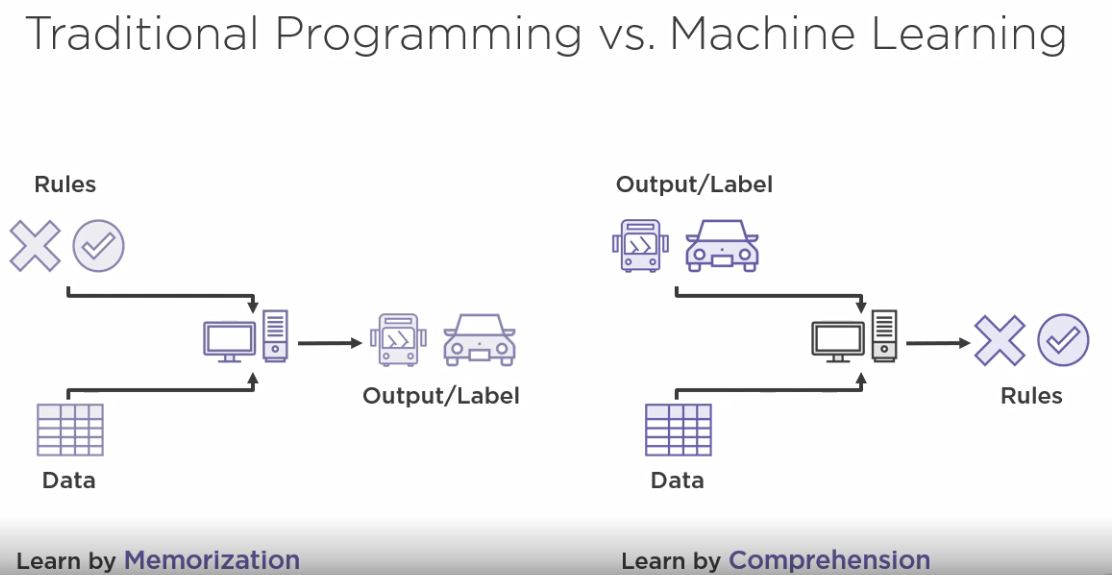
To start with a quote…’what is Machine Learning?’

*Figure: Reproduced Quote Image from Pluralsight (Kurata, 2016)*



The difference between Machine Learning and ‘traditional programming’ can be illustrated briefly as follows.

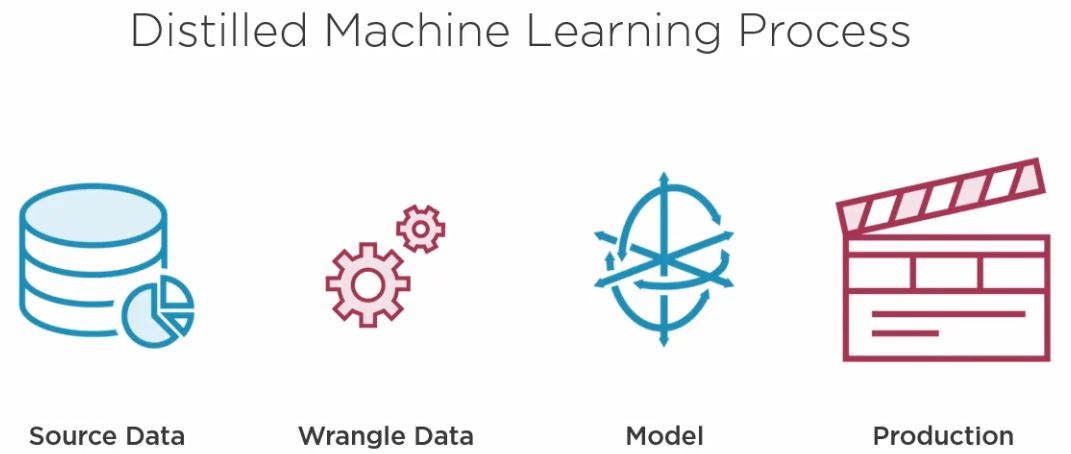
Figure: Traditional Programming v Machine Learning - *Reproduced Image from Pluralsight* (Rhodes, 2020)



This project aims to create a model that can take unseen data and determine a prediction as to whether the transaction is fraudulent, as opposed to a approach such as writing code ti implement a sequence of pre-defined rules.

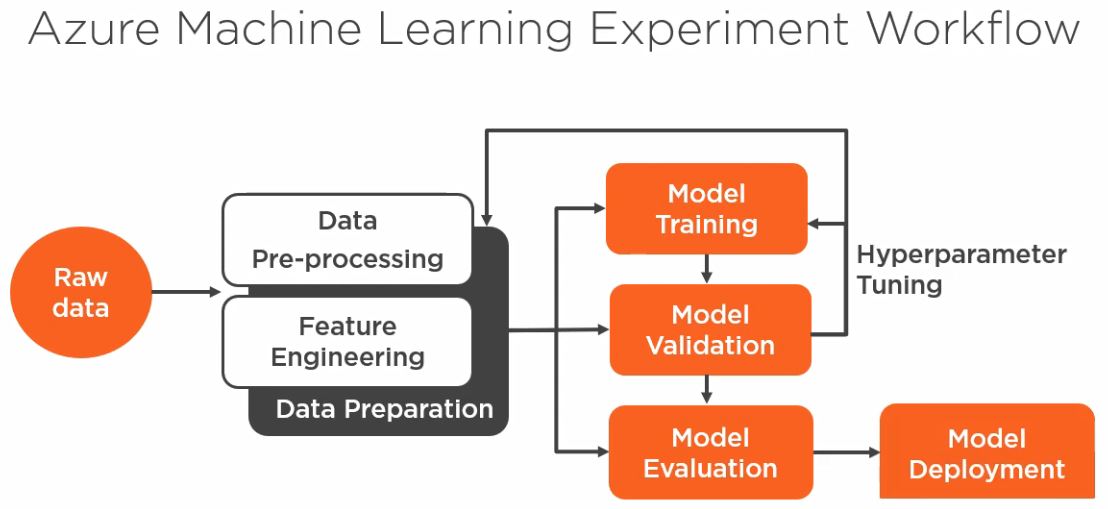
This is a simplified diagram of how the Machine Learning process is applied.

*Figure: Reproduced Image from Pluralsight (Kurata, 2016)*

**

The following figure shows the steps in Azure Machine Learning Studio about which I will provide further implementation details in Section 4.2 through to Section 4.5.

Figure: *Reproduced Image from Pluralsight* (Rhodes, 2020)



## Credit Card Fraud – The Azure Workspace/Machine Learning Studio

The steps to create an Azure account and Workspace are well documented by Microsoft, and I have not sought to reproduce them in detail in this document.

Similarly, the set up required for Azure Machine Learning Studio is equally well documented and accessible from within the Azure portal.

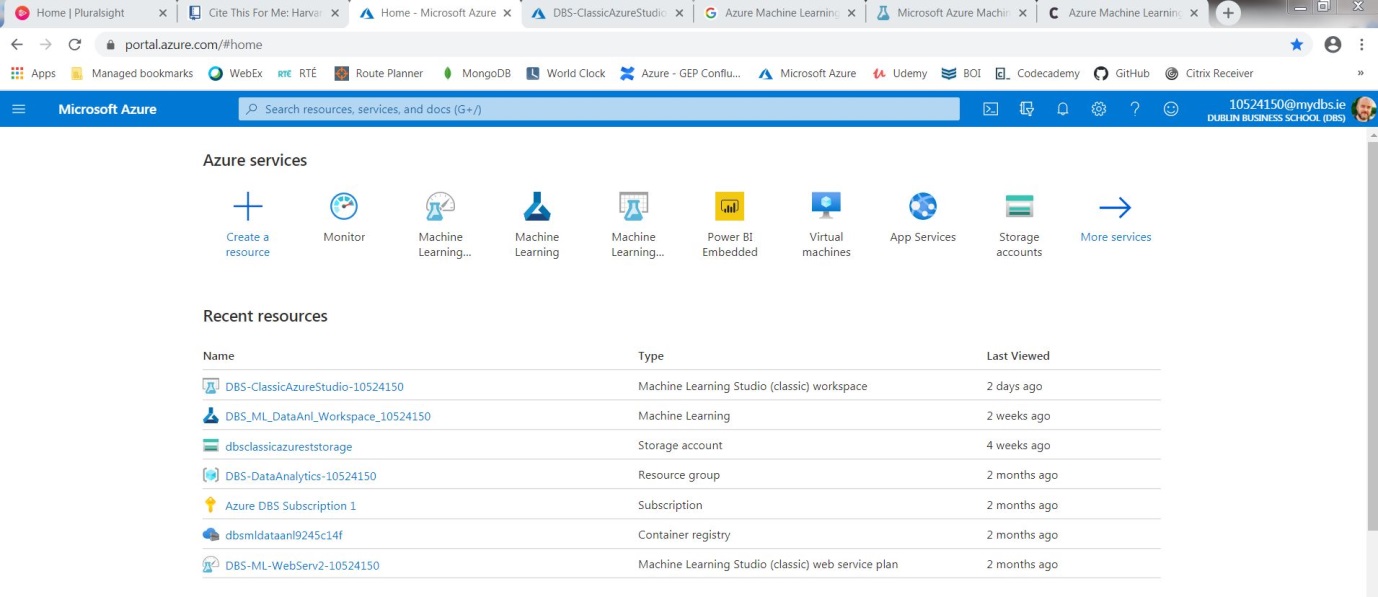
In brief; a description of Azure Workspaces can be found here; <https://docs.microsoft.com/en-us/azure/machine-learning/concept-workspace>

A description of the Azure Machine Learning Studio/Services offering is described here;

<https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-machine-learning-studio>

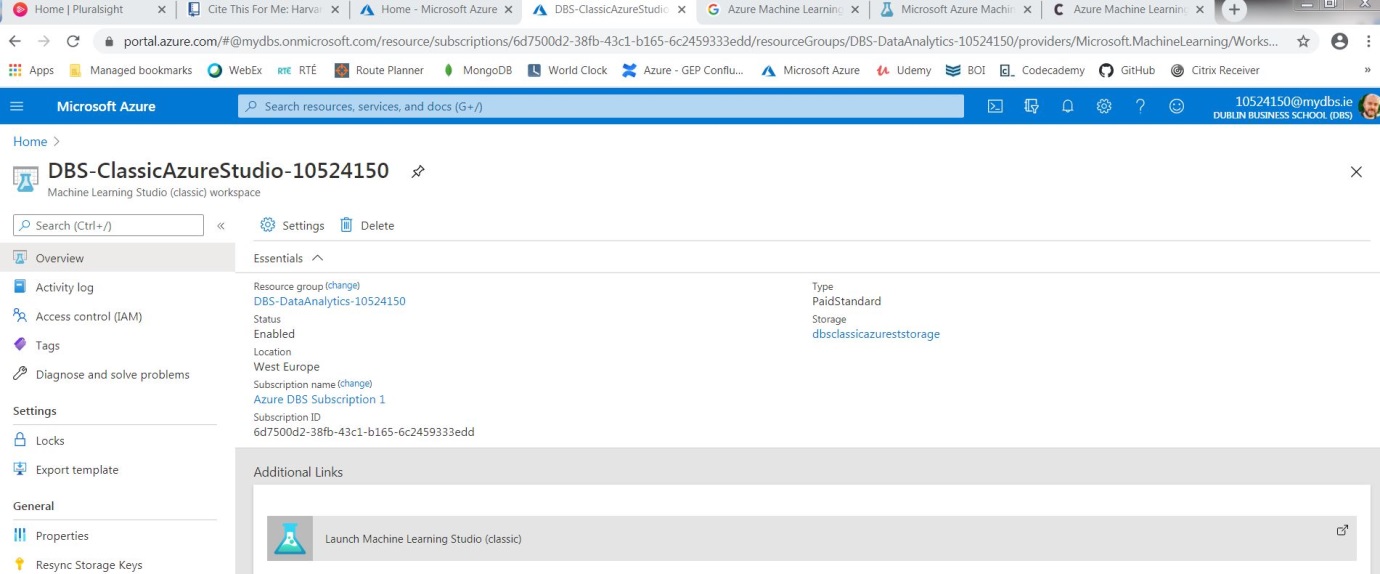
To access the Azure Machine Learning Studio (Classic) where I developed my project the first step is to log onto the Azure Portal, which I set up with my DBS account.

*Figure: Azure Portal (my DBS account)*



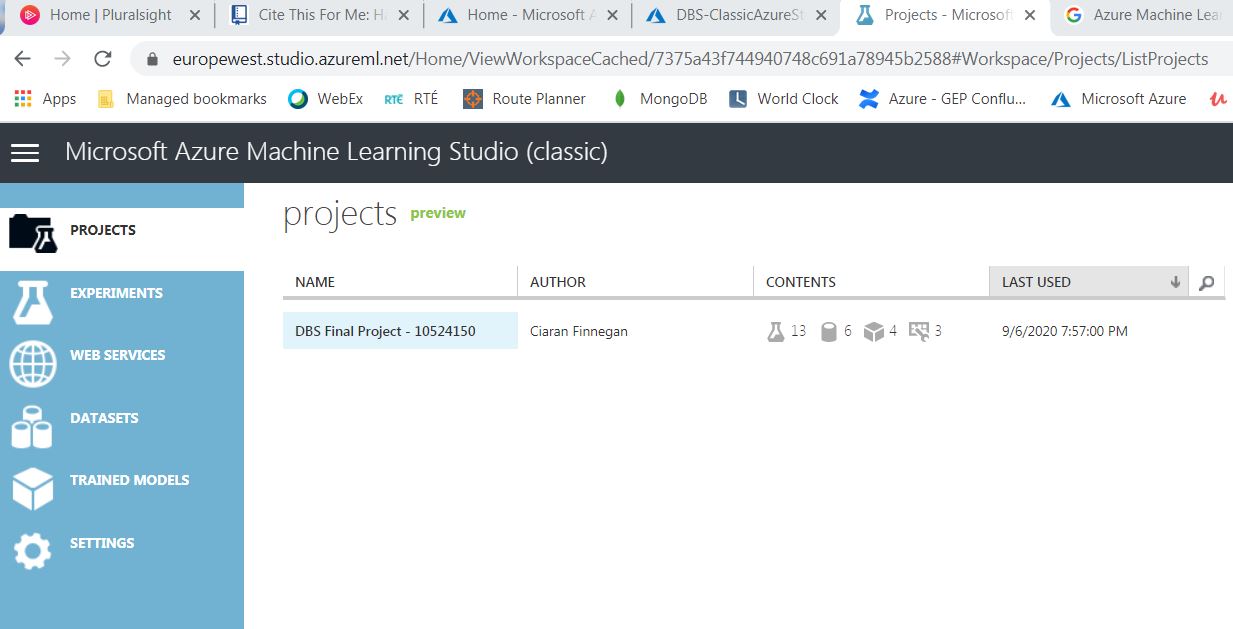
A workspace has been created by me for a Machine Learning Studio (Classic) service.

Figure: Azure ML Studio (Classic) Workspace



Launching the Machine Learning Studio (Classic) services will, after additional user verification, open the ML Studio (Classic) application itself.

Figure: Microsoft Azure Machine Learning Studio (classic)



This ML Studio follows many of the conventions of similar products on the marketplace in terms of organising work under a ‘Project’ structure.

This is a screenshot of my project work and the Project groups together;

* Datasets
* Experiments

The Experiments use the datasets, or outputs of other experiments, to build up the predictive credit card fraud model for this project.

Once ready, my ‘final’ experiment is promoted to a ‘Web Service’ which can then be invoked externally (by my Shiny R application in the case of this project).

The following sections are a sequential analysis of the experiments used in the ‘Project’ to progress through all the steps of the Machine Learning process.

Experiments have been numbered in sequence but the machine learning process has iterated backwards and forwards across the experiments as refinements and alternative options were identified.

**Note: Why use the ‘Classic’ version of the Microsoft Machine Learning Studio?**

In his brief article from 2019 on Codit, Sriram Narayanan***n***, describes the additional features that the more modern Microsoft ML Services offers in comparison to the ‘classic’ studio. Microsoft itself tries to encourage use of the more modern ML ‘Services’ interface.

Working iteratively through the prototype phase of this project, I determined that the ‘classis’ studio was a better option for this delivery for the following reasons;

* **Cost**. The Azure charge for the ‘classic’ studio is very low, and includes the deployment of Web Services / Endpoints. Azure Machine Learning Services is significantly more costly for deploying REST Endpoints on AKS Clusters.
* **Complexity and maturity**. Some of the deployment aspects of Microsoft Azure Machine Learning Services are still in ‘preview’ mode. During Prototype development I had to re-code errors within the Python in certain Jyputer Notebooks when using ML Services examples. I believe that the ‘classic’ option was a more robust platform on which to develop a full ‘end-to-end’ solution.
* **Training**. The Pluralsight courses, to which I had access, had a greater range of training material on ‘classic’ and were an important reference tool for me on this project.

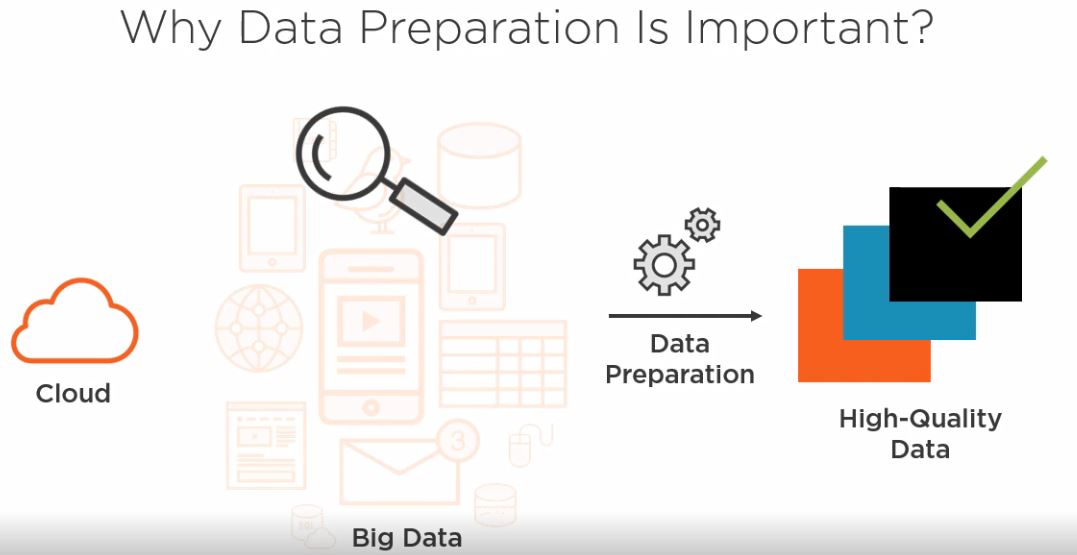
*Figure: Overview of Azure ML Studio (classic) environment for this project – Experiment View*

A screenshot of a computer screen

Description automatically generated

## Credit Card Fraud Dataset – Analysis and Preparation

Figure: *Reproduced Image from Pluralsight*  (Srinivasulu, 2019)



Section 2 of this document described the importance of ‘Feature Engineering’ in then general creation of a credit card fraud predictive model.

To focus specifically on my dataset, feature engineering was important because:

* *My original credit card dataset has 380 columns*. Almost certainly, only a fraction of these columns contains information that will directly influence the accuracy of the final model. It will be necessary to identity those columns that build the most accuracy and performant predictive model for credit card fraud.
* *The dataset is effectively ‘clean’ but still needs to be checked for ‘invalid’ data*. There are no invalid characters in the dataset rows, but missing or useless data needs to be identified, if present.
* *40 columns in the original dataset are non-numeric features and will need some form of re-coding*. Many machine learning algorithms can process non-numeric features, but accuracy is likely to be improved if String features are manipulated before the modelling process begins.

This section of the document details the set up and execution of the following experiments:

* Experiment 1: Data Cleansing
* Experiment 2: Feature Engineering
* Experiment 3: Feature Selection

Exploratory Data Analysis (EDA) is carried out throughout these experiments but the Shiny App UI provides useful graphical descriptions of the dataset. This can be seen in Section 5.1 of this document.

### Experiment 1: Data Cleansing

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged for the data cleansing routines.

*Figure: Experiment 1: Data Cleansing*

A screenshot of a cell phone

Description automatically generated

Appendix 9.3 of this document details the specific steps in this experiment.

The result of this experiment can be summarised as:

* Dataset reduced to 250 columns of potentially ‘useful’ data.
* Top 5% of outlier values in transaction amount ‘clipped’ to reduce distortion in modelling process.
* Generation of an interim dataset for use in Experiment 2.

### Experiment 2: Feature Engineering

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged for the feature engineering routines.

*Figure: Experiment 2: Feature Engineering*

A screenshot of a cell phone

Description automatically generated

Appendix 9.3 of this document details the specific steps in this experiment.

The result of this experiment can be summarised as:

* Conversion of String datatypes to ‘Categorical’ features
* Grouping of Country Code categorical data and numerical encoding of all categorical features.
* Balancing of dataset (via R code routine) to a 50/50 Fraud/Non-Fraud split.
* Identification and removal of a sub-set of highly correlated features.
* Generation of another interim dataset, which will be the input for Experiment 3.

### Experiment 3: Feature Selection

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged for the feature selection routines.

*Figure: Experiment 3: Feature Selection*

A close up of a map

Description automatically generated

Appendix 9.3 of this document details the specific steps in this experiment.

The result of this experiment can be summarised as:

* Taking the output of the feature engineering steps in Experiment 1 + 2 and generating a predictive fraud model.
* Obtaining a list of features scored in order of importance to the predictive model. The ‘Permutation Feature Importance’ module produces this output.

This experiment was run multiple times with various modelling algorithms, based on comparisons seen in later experiments. The ‘Two-Class Logistic Regression’ algorithm provided the best performing and accurate model and was, hence, used to determine the final list of parameters selected for the model.

This choice of features has a direct impact on the feature set captured in the Shiny App UI and passed to the Rest Endpoint for the predictive model.

## Credit Card Fraud – Building the Azure Model

After a series of iterations backwards and forwards through the experiment sequences, I believed that I now had a refined credit card dataset with which I could run a final batch of modelling experiments.

*Figure: Representation of ML Modelling Process Reproduced from Edureka (Lateef, 2020)*

A close up of a map

Description automatically generated

The pattern of operations followed the illustration above, but my primary objectives were:

* Determine which classification algorithm, which is available for use in Azure Machine Learning Studio (classic), would be most effective for generating a predictive fraud model based on my credit card transaction dataset. Criteria for algorithm selection would be:
  + Accuracy Score
  + Recall – how well actual Fraud is detected
  + Performance
* Demonstrate the impact on fraud prediction model accuracy, and other metrics, introduced by the following modelling actions:
  + Feature Engineering
  + Cross Validation
  + Hyperparameter tuning

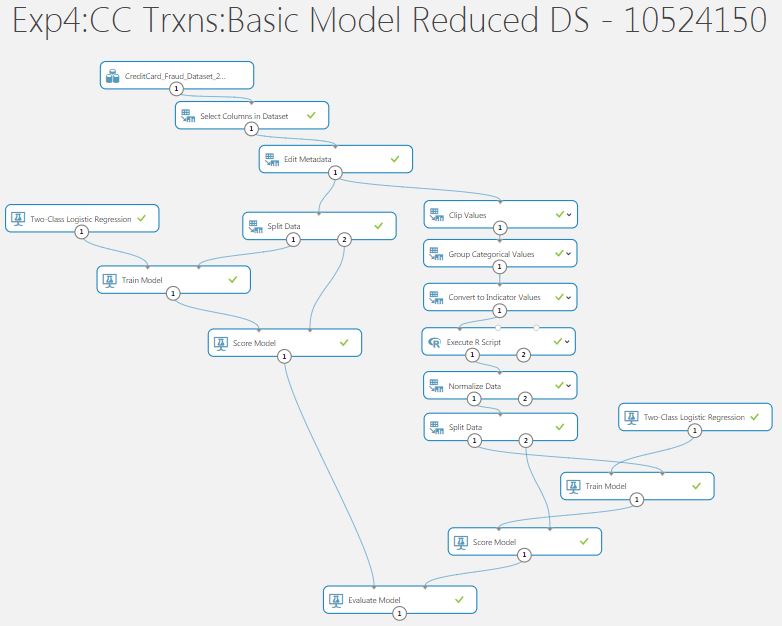
This section of the document details the set-up and execution of the following experiments:

* Experiment 4: Basic Model Evaluation with Feature Engineering
* Experiment 5: Model Evaluation using Cross Validation and Hyperparameter tuning
* Experiment 6 + 7: Comparison of Multiple Classification Algorithms

### Experiment 4: Basic Model Evaluation with Feature Engineering

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged to assess the benefits of Feature Engineering.

*Figure: Experiment 4: Feature Engineering and Model Evaluation*



Appendix 9.3 of this document details the specific steps in the left hand side (LHS) and right hand side (RHS) of this experiment as they largely replicate the work in Experiments 2 and 3.

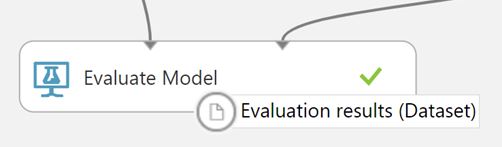
The result of this experiment can be summarised as:

* A demonstration of the impact of feature engineering on model accuracy and other metrics.
* Possible trade-offs that might be acceptable in the modelling process.

Again, ‘Two-Class Logistic Regression’ is used because of evaluation results in later experiments feeding back into this ‘final’ version of Experiment 4.

*Model Evaluations*

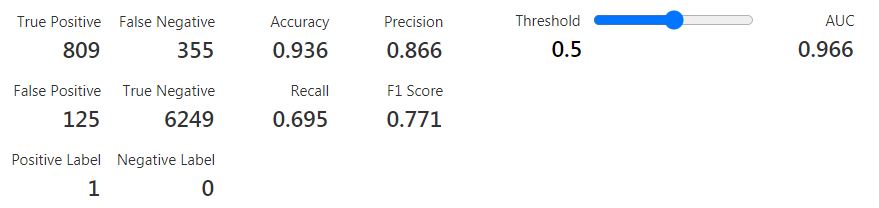
The ‘Evaluate Model’ module provides takes two inputs and provides the key scoring metrics on comparative models as an output.



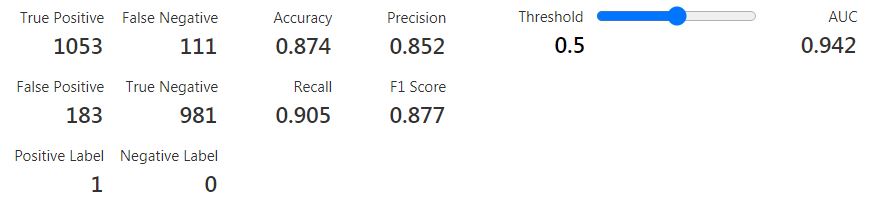
In Experiment 4:

* The ‘*Scored dataset*’ was the model generated without Feature Engineering, except for the conversion of String features into Categorical features.
* The ‘*Score dataset to compare*’ was the model generated with the Feature Engineering routines in Experiments 1, 2, and 3.

The ‘*Scored dataset*’ produced the following scores:



The ‘*Scored dataset to compare*’ produced the following scores:



*Model Score Assessments*

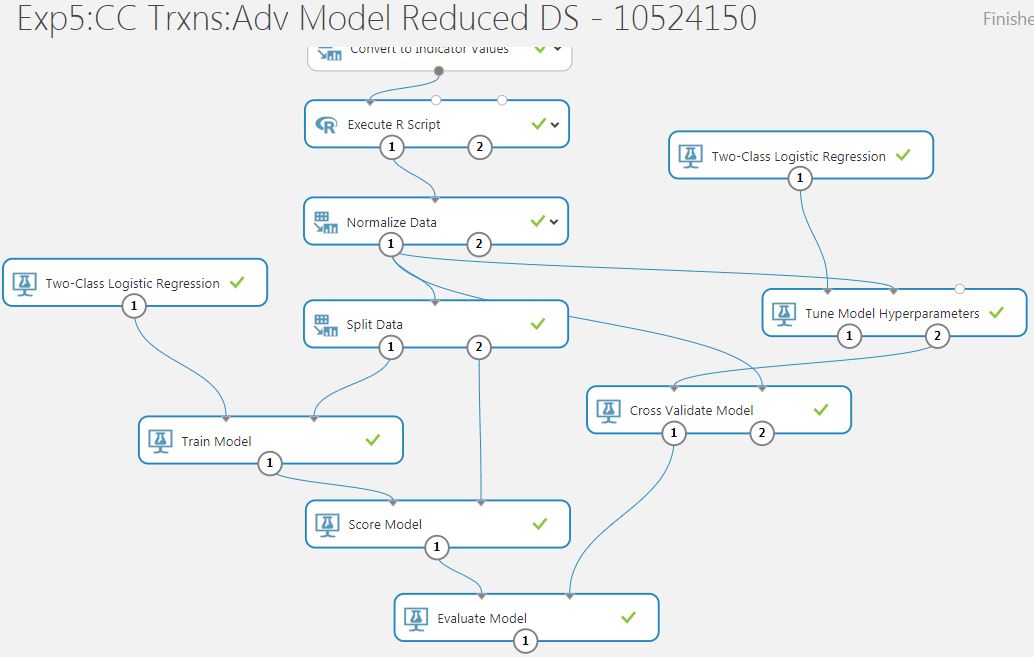
Feature Engineering does not improve the overall accuracy of my credit card predictive model for fraud, but it is much better at detecting actual fraud cases (higher recall value).

### Experiment 5: Model Evaluation with Cross Validation/Hyperparameter Tuning

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged to assess the benefits of using Cross Validation and hyperparameter tuning.

*Figure: Experiment 5: Cross Validation and Hyperparameter Tuning*

*Note:- This image has been deliberate truncated to focus on the modules after Feature Engineering.*



Appendix 9.3 of this document details the specific configurations of the ‘Tune Model Hyperparameter’ and ‘Cross Validate Model’ modules.

The result of this experiment can be summarised as:

* Experiment 4 conducted a straightforward Test/Train split of the dataset for modelling. Can we determine if Cross Validation will improve the reliability of my predictive model for credit card fraud detection?
* Azure Machine Learning Studio (classic) allows for an automated process to tune the hyperparameter values on an algorithm. Does this also contribute to better fraud prediction for my dataset?

Again, ‘Two-Class Logistic Regression’ is used because of evaluation results in later experiments feeding back into this ‘final’ version of Experiment 5.

*Model Evaluations*

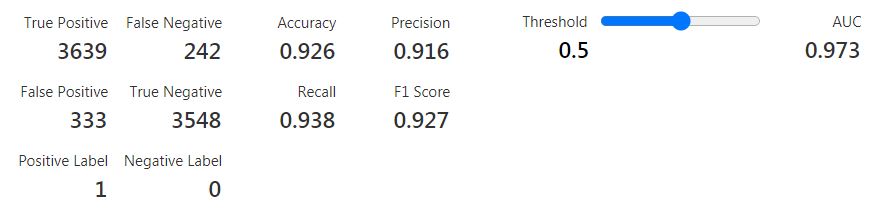
As before, the ‘Evaluate Model’ module provides takes two inputs and provides the key scoring metrics on comparative models as an output.

In Experiment 5:

* The ‘*Scored dataset*’ was the model generated with Feature Engineering in Experiment 4.
* The ‘*Score dataset to compare*’ was the model generated using Cross Validation on the dataset and tuned hyperparameters for the Two-Class Logistic Regression algorithm.

The ‘*Scored dataset’* is unchanged from Experiment 4.

The ‘*Scored dataset to compare*’ produced the following scores:



*Model Score Assessments*

Using Cross Validation and hyperparameter tuning in Experiment 5 has produced a model that scores almost as well in ‘Accuracy’ as the LHS model Experiment 4 (0.936 vs 0.926).

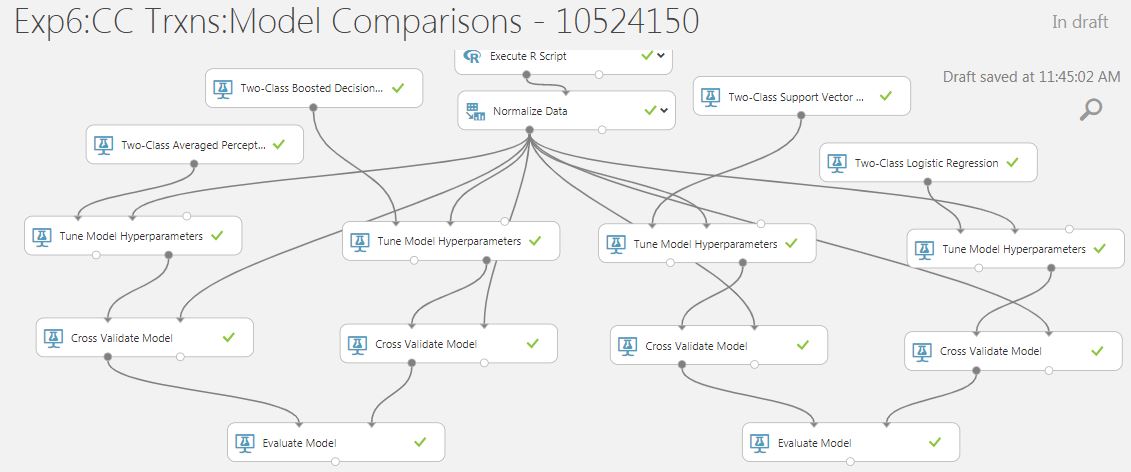
However, the ‘Recall’ score in Experiment 5 is higher again (0.938) and is thus even better at finding Fraud that either of the models in Experiment 5.

### Experiment 6: Comparison of Multiple Classification Algorithms (1)

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged to assess the performance of multiple Classification algorithms.

*Figure: Experiment 6: Comparing Classification Algorithms*

*Note:- This image has been deliberate truncated to focus on the modules after Feature Engineering.*



Based on the results from Experiment 5, Cross Validation and hyperparameter tuning will be applied to all models built in further experiments to create my credit card predictive model for fraud detection.

The result of this experiment can be summarised as:

* Compare results of four similar ‘Two-Class’ Classification algorithms when creating a predictive model based on my credit card fraud dataset. The algorithms being compared in this experiment are:
  + Two-Class Averaged Perceptron.
  + Two-Class Boosted Decision Tree.
  + Two-Class Support Vector Machine.
  + Two-Class Logistic Regression.

The selection of classification algorithms in the Azure Machine Learning Studio (classic) is limited to nine options, of which I choose eight. The other classification algorithms specialise in multi-class problems.

*Model Evaluations and Assessments*

Appendix 9.3 of this document provides a breakdown of all Experiment 6 and 7 evaluation scores for each model.

**‘Two-Class Logistic Regression’ performs best, based on a combination of ‘Accuracy’ and ‘Recall’.**

Other observations on the algorithm performances were (based on a 25K row dataset with 39 features):

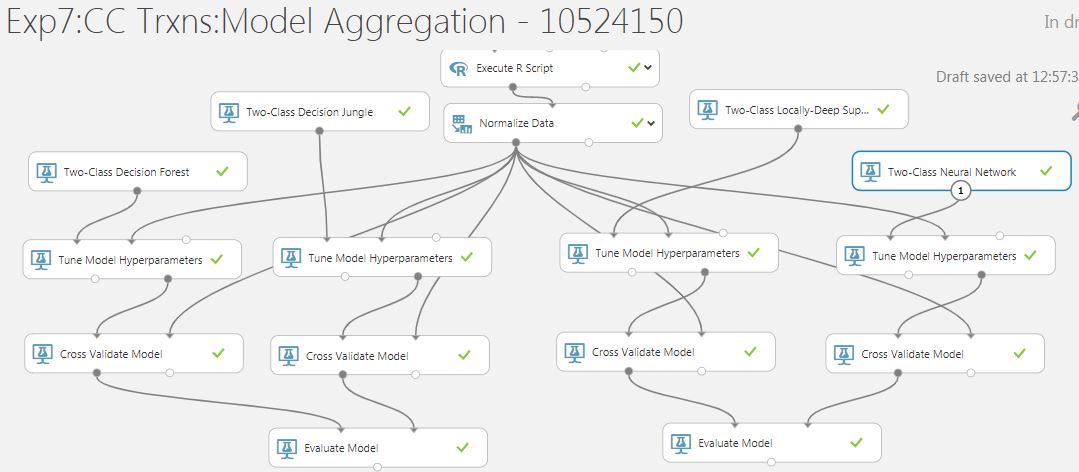
* The Two-Class Averaged Perceptron algorithm was the quickest to run (< 1 minute) and complete. The Microsoft documentation describes this as a simplified version of a neural network. It is sometimes favoured when the goal is speed over accuracy. (Microsoft, 2019).
* The Two-Class Boosted Decision Tree took the longest to run and complete. Hyperparameter tuning alone took 10+ minutes, and the model was not available for scoring for nearly 20 minutes. The Azure Machine Learning Studio (classic) contained a tutorial recommending this algorithm for client credit risk solution, but performance with my dataset was a concern. (Normalization was probably a redundant step in this modelling process but was left in place for simplicity.)
* The Two-Class Support Vector algorithm took 5+ minutes to complete the modelling process. (The second longest). Microsoft documentation recommends this for simpler datasets where the aim is, again, speed over accuracy. Results were good but performance was slow.
* Two-Class Logistic Regression was dependent on conversion of non-numeric features but performed the best overall.

### Experiment 7: Comparison of Multiple Classification Algorithms (2)

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged to assess the performance of further multiple Classification algorithms.

*Figure: Experiment 7: Comparing Classification Algorithms – Pt2*

*Note:- This image has been deliberate truncated to focus on the modules after Feature Engineering.*



The result of this experiment can be summarised as:

* Compare results of four similar ‘Two-Class’ Classification algorithms when creating a predictive model based on my credit card fraud dataset. (). The algorithms being compared in this experiment are:
  + Two-Class Decision Forest.
  + Two-Class Decision Jungle.
  + Two-Class Locally-Deep Support Vector Machine.
  + Two-Class Neural Network.

These are possibly more complex algorithms with greater processing overhead and are included in the project to compare with the group of algorithms in Experiment 6.

*Model Evaluations and Assessments*

Appendix 9.3 of this document provides a breakdown of all Experiment 6 and 7 evaluation scores for each model.

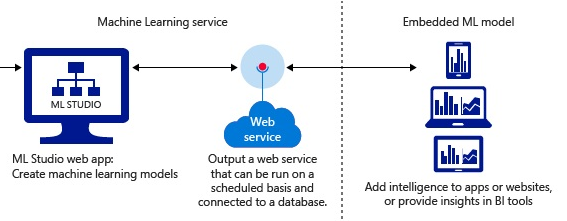
None of the Experiment 7 algorithms generated superior results, in terms of ‘Accuracy’ and ‘Recall’ when compared to the Two-Class Logistic Regression based model.

## Credit Card Fraud – Deploying the Azure Model

My iterations through the experiments to evaluate the best algorithm, including the optimum training process, provided me a training model that I now wanted to deploy into production.

This would allow me to host the model in Azure and invoke that model to display fraud predictions on ‘new’ credit card transactions.

*Figure: Reproduced from MicroStrategy Community*(Sonobe, 2017)



My objectives, at this stage of the project lifecycle, were to:

* Prepare and validate a ‘final’ model based on my refined feature engineering routines and training process, using my chosen classification algorithm.
* Create a ‘Predictive’ version of the trained model, in preparation for the set-up or a Web Service to be hosted in Azure. Deploy this credit card predictive fraud model as a Web Service hosted in Azure and test the deployment with sample data.
* Update the Shiny App UI code with R code that invokes the API to return a real time predictive score on the likelihood of Fraud for a given new credit card transaction, selected by the user through the project UI. (See Section 5 of this document for details on the code routines to extract key data elements from ‘new’ card transactions and pass them to the API for the fraud model).

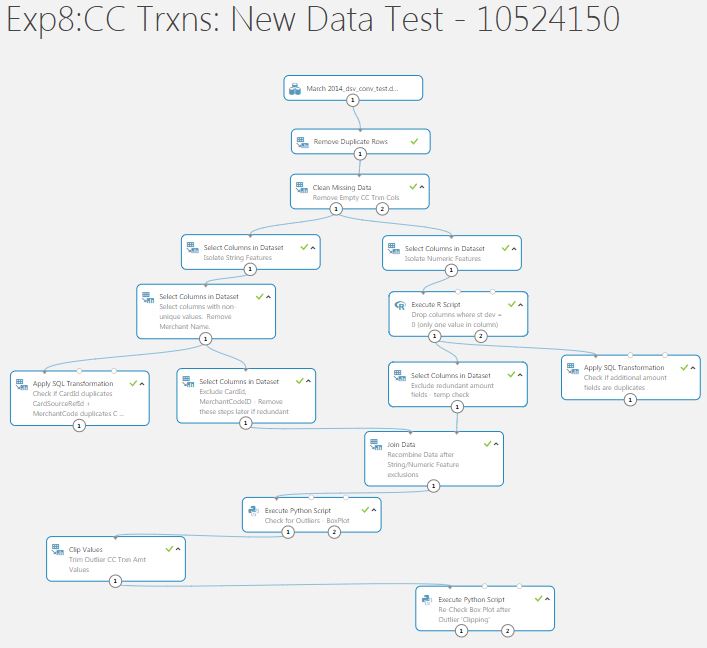
This section of the document details the set-up and execution of the following experiments/actions:

* Experiment 8: Repeat of initial Feature Engineering routines with larger dataset.
* Experiment 9: Creation of ‘Predictive’ model.
* Deployment and validation of Web Service for trained model.

### Experiment 8: Feature Engineering on Larger Dataset

The figure below illustrates how the Azure Machine Learning Studio (classic) modules were arranged to assess process initial Feature Engineering routines on the larger credit card transaction dataset.

*Figure: Experiment 8: Feature Engineering on Larger Dataset*

**

Experiment 8 is a re-execution of Experiment 1 but on the larger 25K dataset.

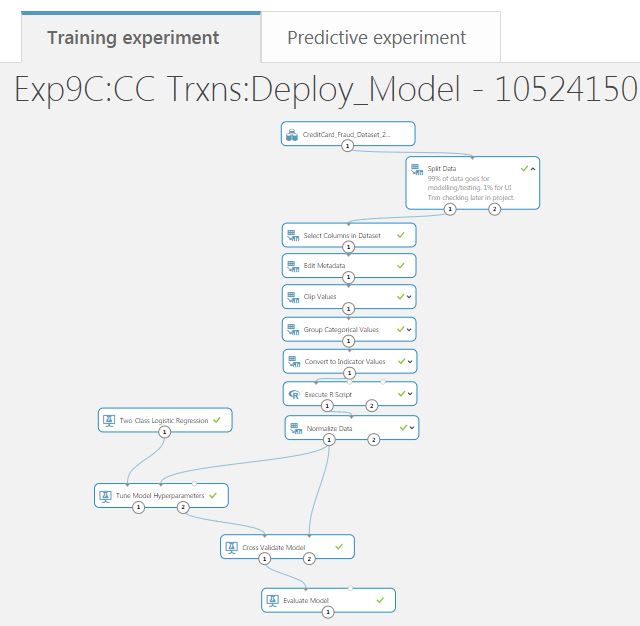
The larger dataset is being introduced at this point in the project to provide a greater volume of data for the training process, and thus ideally increase the reliability of the predictive fraud model.

### Experiment 9: Creation of Predictive Fraud Model for Deployment

Experiment 9 is drawn from the results and conclusions from earlier experiments.

The ‘Training Experiment’ in the illustration below represents the ‘final’ model creation approach.

*Figure: Experiment 9 : Training Model*



(The ‘9C’ numbering convention is the result of a number of iterations on this experiment).

The purpose of this Experiment is to create a training model which will then be converted into a ‘Predictive’ model.

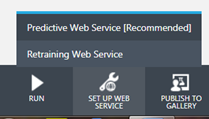
The ‘Predictive’ model is the basis for the deployment of a Web Service to allow external access (from my Shiny R application) to the scoring model for credit card fraud.

The experiment above is under the ‘Training’ tab.

*Generating a Predictive Experiment*

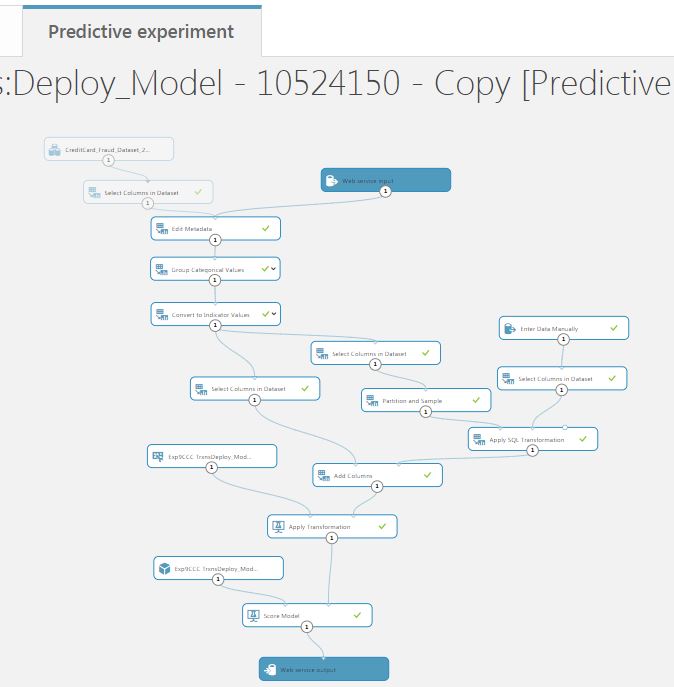
The Azure Machine Learning Studio (classic) provides an option for any experiment with a trained model to be deployed as a Web Service.

*Figure: Option to generate Predictive experiment*



This creates a ‘stripped down’ version of the Training experiment called the ‘Predictive experiment’.

*Figure: Experiment 9: Predictive experiment*

**

*Web Service Inputs / Outputs*

As some of the remaining modules will be redundant for the real time ‘one-by-one’ scoring of transactions, which is a key requirement of my credit card fraud prediction project, I have removed other elements of the ‘Training experiment’ that were brought across.

Key features of the Predictive experiment, as show in the illustration, are:

* Azure Machine Learning Studio (classic) introduces Web Service input and output modules. These determine the interface points to the model in deployment.
* I moved the Web Service input to a point after the Feature Selection module. This is done so that the API code in my Shiny R application will only need to pass the sub-set of features directly required by the module, and not the much larger (post Feature Engineering) dataset.
* I have created additional modules to generate a manual one-hot encoding. This is required because tests failed during verification of the deployed model when transactions were being processed ‘one-by-one’. Single credit card transactions would not generate the additional non-numeric features created during the modelling process with the larger dataset.

One validated the Predictive experiment can be deployed as a Web Services, hosted within Azure.

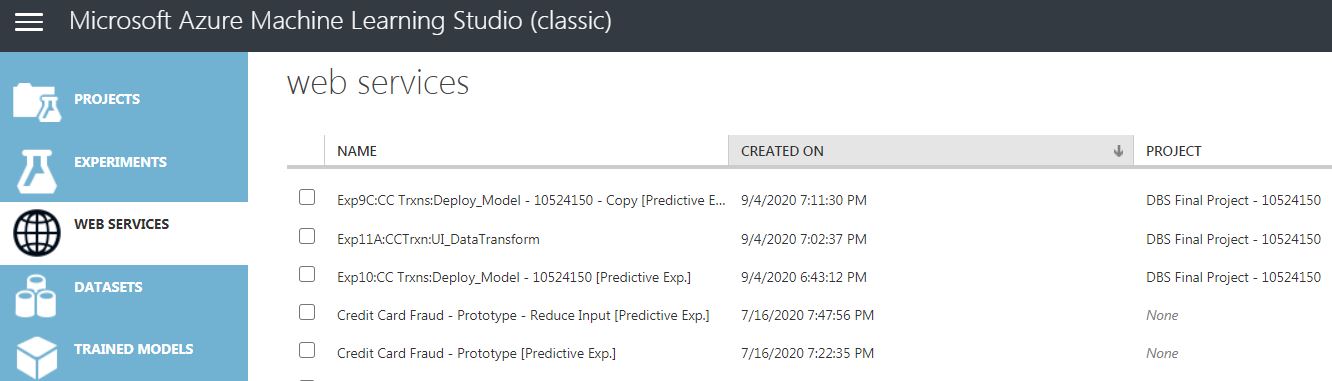
*Figure: Option to Deploy Web Service*



### Deployment and Validation of Web Service for Predictive Fraud Model

Azure Machine Learning Studio (classic) maintains a list of generated Web Services, which can be accessed through the Studio interface.

*Figure: Web Services List*

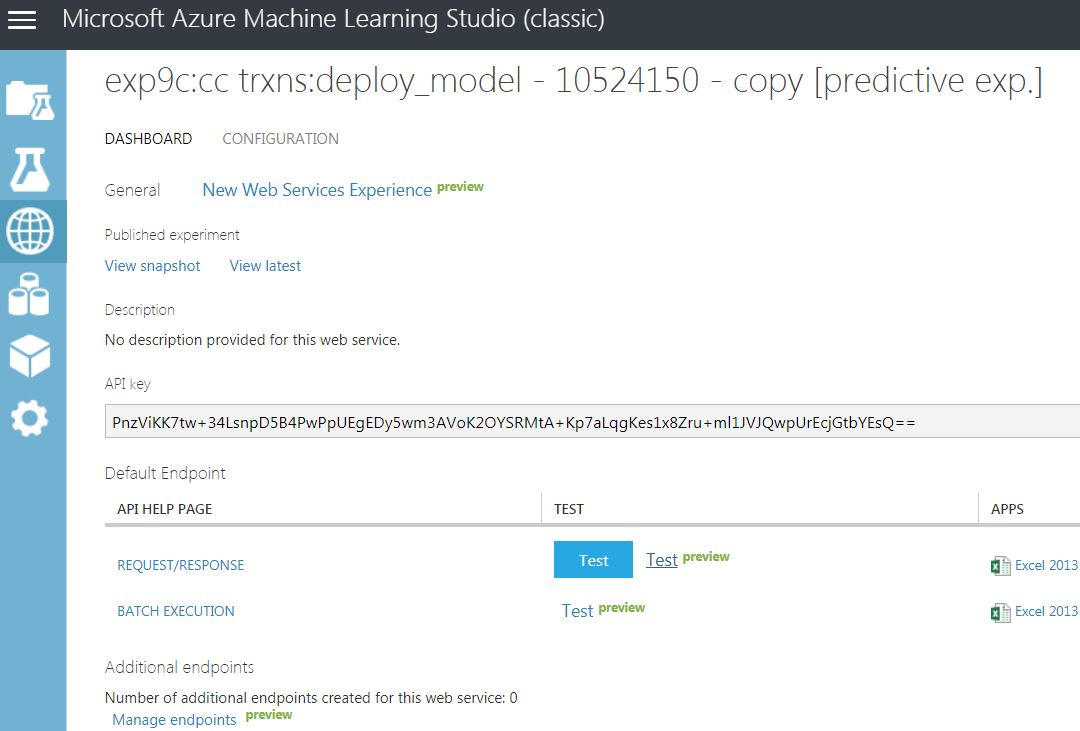


The Web Services that I have generated in my project as part of ongoing research, for the Interim Prototype, and for the final predictive credit card fraud model can be seen in this illustration above.

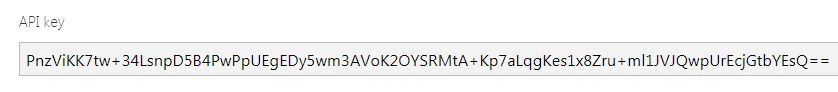
*How to validate the Web Service through Azure?*

The generation process for a Web Services, if successful, brings the user to a dashboard screen. The illustration below shows the dashboard screen for my final production model for credit card fraud prediction.

*Figure: Web Services Dashboard*



Section 5 of this document will explain more about how this Web Service is consumed but a key element on this page is the API key through which the model can be invoked externally.

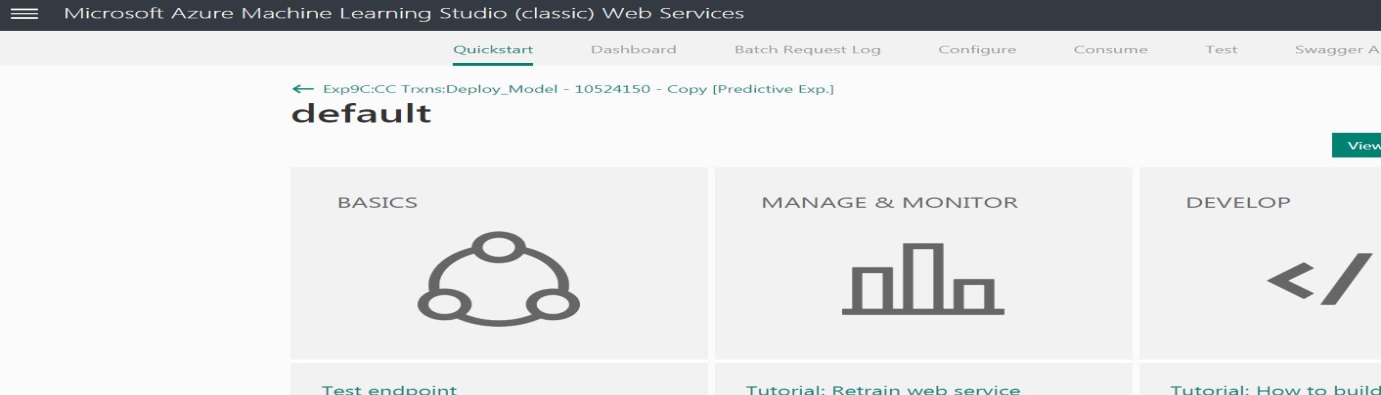


Before I began writing R code to access the API for the fraud model, I needed to verify that the Web Service was working as expected and returning a score for predicting fraud on my credit card transaction.

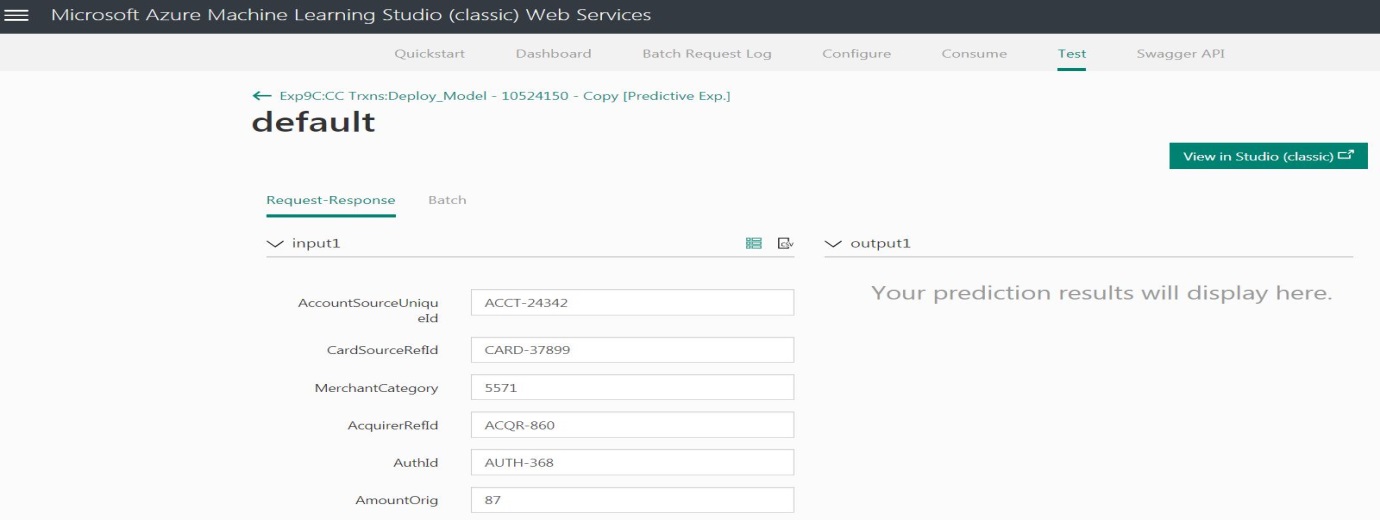
Azure Machine Learning Studio provides a separate Web Services portal to test and manage these hosted endpoints.

Using this portal, shown in the illustrations below, I was able to validate my credit card fraud predictive model was working correctly.

*Figure: Azure ML Studio Web Services Portal – Main Screen*



*Figure: Azure ML Studio Web Services Portal – Test Screen*



Test results return a ‘Score Label’ – ‘1’ for Fraud, ‘0’ for Non-Fraud. A Scored Probability value is also returned, which is a number between 0 and 1 (> 0.5 = Fraud, <0.5 = Non-Fraud).

# Project Implementation (2) – Shiny R Dashboard UI

## Data Visualisations in a Shiny Dashboard

The prototype for this project application..

The..

<images>

## Credit Card Fraud – UI to Check Fraud Predictions

The..

<images>

## Shiny UI – Hosted Application

The..

<images>

# Testing and Results

## User Story ‘Demos’ – Test Results and ‘Feedback’



### User Story 4: Initial Data Modelling – Review and Evaluation

Goal: *Build a basic credit card fraud predictive model in Azure ML Studio (classic) based on a small subset of transactions dataset*.

Assessment of robustness of code and functionality delivered:

1. Goal Achieved – August 1st 2020. User Stories 1 – 3 provided enough research and background to set up Azure ML workspace for ML Studio (classic).

*Figure: User Story 4 demonstration*

A screenshot of a cell phone

Description automatically generated

1. Model generated with manual selection of features and elementary. Tests with ‘Evaluate Model’ module displayed Accuracy results of ~82%. ‘Recall’ value extremely poor but model acceptable for prototype.

*Figure: User Story 4 Test Model Results*

*A screenshot of a cell phone

Description automatically generated*

### User Story 5: Basic Shiny App – Review and Evaluation

Goal: *Build a basic Shiny R dashboard app that displays basic EDA of my credit card dataset and has a placeholder screen for fraud detection interface.*

Assessment of robustness of code and functionality delivered:

1. Goal Achieved – August 7th, 2020.

*Figure: User Story 5 demonstration*

*A screenshot of a cell phone

Description automatically generated*

1. Quick Turnaround from User Story 1. Basic Shiny Dashboard App running without error from within RStudio environment.

### User Story 6: Integrated Prototype – Review and Evaluation

Goal: *Add R code to R Shiny Dashboard to invoke basic card fraud model with fixed data inputs. Host working Shiny App online.*

Assessment of robustness of code and functionality delivered:

1. Goal Achieved – August 14th, 2020. This working prototype was released online with a basic user guide as part of the Interim Report for the project.

*Figure: User Story 6 demonstration*

A screenshot of a cell phone

Description automatically generated

### User Story 7: Enhanced Modelling – Review and Evaluation

Goal: *Refine credit card model with full ML workflow processes. Enhance UI to select ad-hoc credit card transactions.*

Assessment of robustness of code and functionality delivered:

1. Goal Partially Achieved – September 5th, 2020.
2. Full end-to-end ML workflow applied to create a production ready model for credit card fraud prediction. Tested and validated in the Azure ML Studio Web Services portal.

*Figure: Web Services Portal Testing of ‘final’ predictive model.*

A screenshot of a cell phone

Description automatically generated

1. Shiny App UI only partially updated. Complexity of rebuilding model left no time to complete this section of the User Story. The UI is reading in new fixed files but there is no option to select a transaction file at random by the user. Carried over to User Story 8.

### User Story 8: Enhanced UI – Review and Evaluation

Goal: ***(Revised)*** *Redeploy new credit card fraud model in Azure. Update code in Shiny R Dashboard to:*

* *Invoke new API*
* *Allow for ad-hoc selection of ‘new’ transactions to submit to predictive fraud model*
* *Display improved data visualisation graphs on UI based on credit card dataset*

Assessment of robustness of code and functionality delivered:

1. Goal Achieved – September 17th, 2020.

### User Story 9: Presentation Preparation – Review and Evaluation

Goal: *Refine UI in preparation for Final project demonstration.*

Assessment of robustness of code and functionality delivered:

1. Goal Achieved – September 25th, 2020.

## Final Project Assessment

The project is intended to demonstrate …

## Project Plan 2020: Final Status – 25th September 2020

**(Produced using the Team Gantt online portal)**

A screenshot of text

Description automatically generated

# Project Location and User Guide

## Credit Card Fraud Application: Prototype Location

The prototype for this project application is currently hosted on *shinyapps.io* and the UI can be accessed through this URL;

<https://ciaran-finnegan.shinyapps.io/DBS_CCFraudRShinyApp_10524150/>

## Credit Card Fraud Application: User Guide (Final Project)

The prototype is intended to demonstrate progress to date and provide tangible evidence of the end goals of this project.

*Figure: Final Project*

A screenshot of a cell phone

Description automatically generated

A User Guide, in Microsoft PowerPoint format, is embedded with this report, and has also been submitted separately…

# Project Conclusions

## Where Project Goals Achieved?

The User Stories ...

## Future Design/Deployment Considerations

The Interim Report.....

● Any changes from the interim report should be discussed and justified.

● The student should reflect on the learning experiences gained in doing the project and its relevance to on--going progress as a learner and future practising IT professional.

● This section should also provide a starting point for another student to continue the work.

# Appendices

## Azure Generated Code Segments

The Azure Machine Learning Studio auto-generates codes segments in C#, Python 3.6, and R to access both Azure hosted datastores and invoke APIs to Azure hosted Rest Endpoints for deployed models.

Below are examples of code snippets which have been incorporated into my project.

This code segments reads a subset of the Credit Card dataset and uses the output to generate data visualisations in the R Shiny App.

*Figure: R code snippet to read Azure hosted datastore*

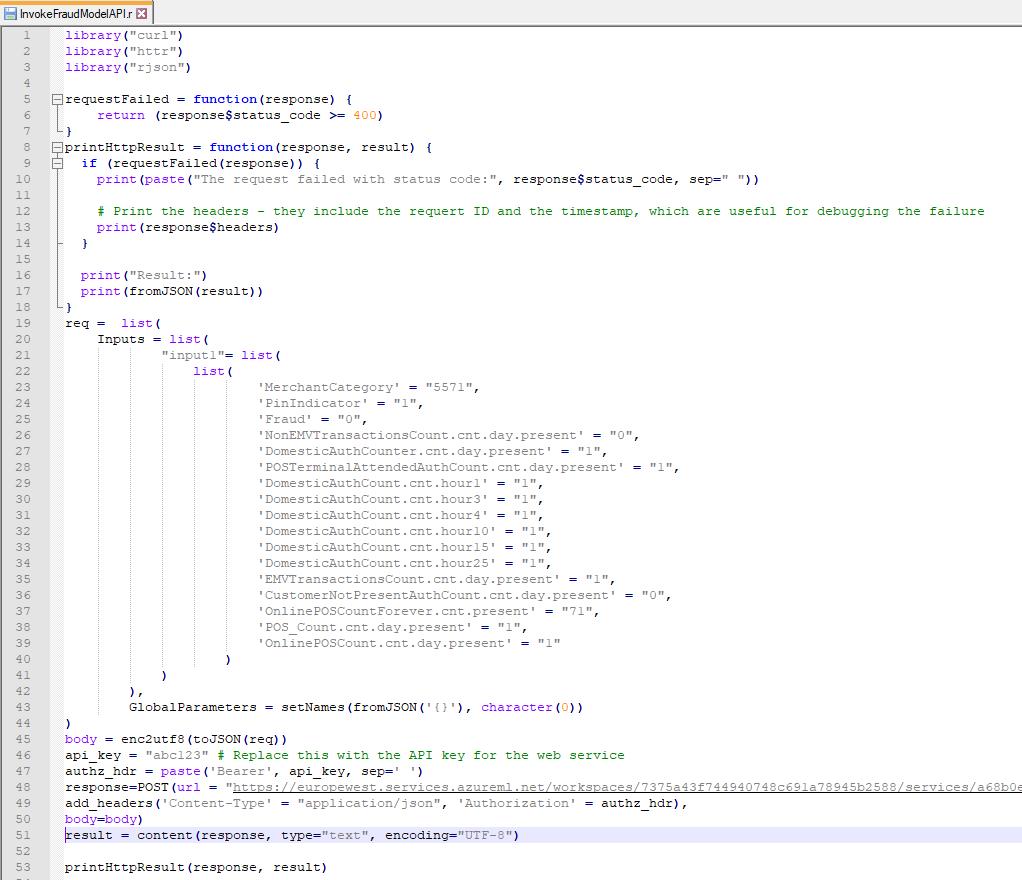
A screenshot of a social media post

Description automatically generated

This code segment invokes the API to pass attributes to the Fraud detection model, hosted in Azure, and returns a prediction score.

(Line 48 has been truncated slightly).

*Figure: R code snippet to read Azure hosted REST Endpoint for Fraud Model*



## Shiny R Application Code Files



### Diagram: The RStudio Cloud Environment

A screenshot of a social media post

Description automatically generated

### The Shiny UI Code

The..

### The R Code Parsing Data and Invoking UI

The..

## Azure Machine Learning Classic Studio Experiments



### Experiment 1: Breakdown

*Breakdown of Experiment*

Exp1: Step 1. Remove duplicate rows. Remove columns with missing data

A screenshot of a cell phone

Description automatically generated

Columns with missing data were seen to have a lot of empty cells. Removal was the best/most straightforward option.

The original dataset started with **380** columns. This transformation reduced the dataset to **362** columns.

A picture containing knife, table

Description automatically generated

### Experiment 2: Breakdown

*Breakdown of Experiment*

### Experiment 3: Breakdown

*Breakdown of Experiment*

### Experiment 4: Breakdown

*Breakdown of Experiment*

### Experiment 5: Breakdown

*Breakdown of Experiment*

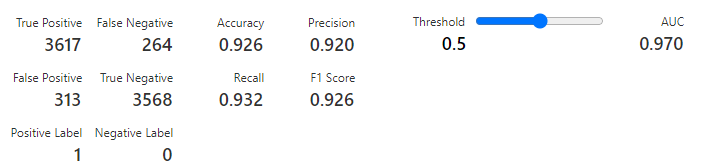
### Experiment 6: Breakdown

*Evaluation results for each classification algorithm*

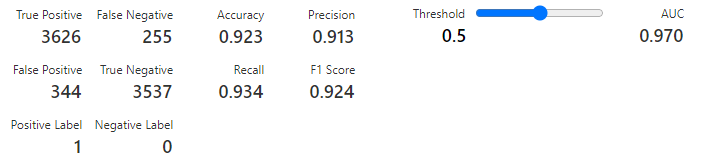
**Exp 6 – LHS**

**Two-Class Averaged Perceptron - v – Two-Class Boosted Decision Tree**

Two-Class Averaged Perceptron



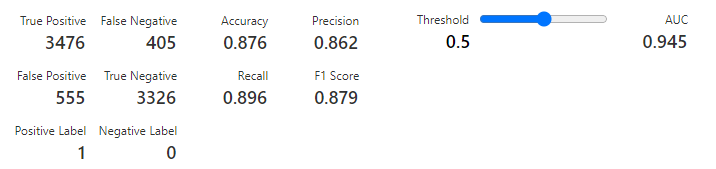
Two-Class Boosted Decision Tree



**Exp 6 – RHS**

**Two-Class Support Vector Machine - v – Two-Class Logistic Regression**

Two-Class Support Vector Machine



Two-Class Logistic Regression



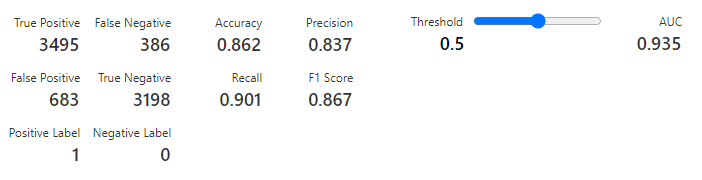
### Experiment 7: Breakdown

*Evaluation results for each classification algorithm*

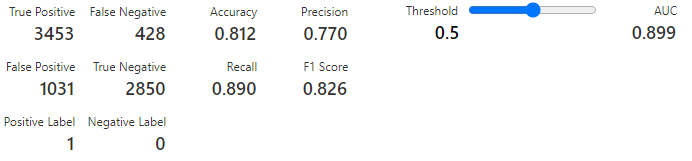
**Exp 7 – LHS**

**Two-Class Decision Forest - v – Two-Class Decision Jungle**

Two-Class Decision Forest



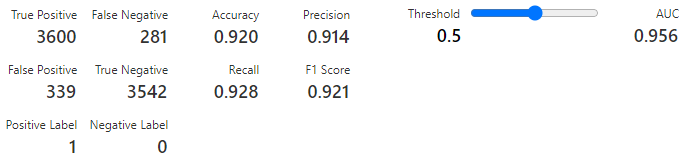
Two-Class Boosted Decision jungle



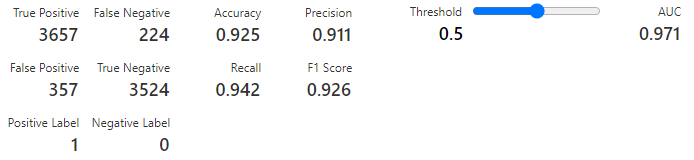
**Exp 7 – RHS**

**Two-Class Locally Deep Support Vector Machine - v – Two-Class Neural Network**

Two-Class Locally Deep Support Vector Machine



Two-Class Neural Network



### Experiment 8: Breakdown

*Breakdown of Experiment*

### Experiment 9: Breakdown

*Breakdown of Experiment*

## Credit Card Fraud Datasets

The datasets..

# References / Bibliography - Interim Report

1Microsoft Tutorial, 2020. *What Is ML Studio (Classic) - Azure*. [online] Docs.microsoft.com. Available at: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio> [Accessed 19 June 2020].

2Krzemiński, D., 2018. *Create Outstanding R Shiny Dashboards With The Semantic.Dashboard Package - Appsilon Data Science | End­ To­ End Data Science Solutions*. [online] Appsilon Data Science | End­ to­ End Data Science Solutions. Available at: <https://appsilon.com/create-outstanding-dashboards-with-the-new-semantic-dashboard-package/> [Accessed 28 June 2020].

3Parsons, T., 2019. *When To Use Waterfall Vs. Agile | Macadamian*. [online] Macadamian. Available at: <https://www.macadamian.com/learn/when-to-use-waterfall-vs-agile/> [Accessed 10 July 2020].

4Stanton, K., 2019. *Card Fraud In Europe - It's Still Increasing..*. [online] Finextra Research. Available at: <https://www.finextra.com/blogposting/16824/card-fraud-in-europe---its-still-increasing> [Accessed 4 August 2020].

5Cherowbrier, J., 2019. *Forecast Of Card Payments Per Day 2006, 2016 And 2026 Kingdom | Statista*. [online] Statista. Available at: <https://www.statista.com/statistics/719708/card-payments-per-day-forecast-united-kingdom/> [Accessed 7 August 2020].

6ICA, n.d. *What Is Customer Due Diligence (CDD)? | ICA*. [online] Int-comp.org. Available at: <https://www.int-comp.org/careers/your-career-in-aml/what-is-customer-due-diligence-cdd/> [Accessed 7 August 2020].

7Machine Learning Group - ULB, 2018. *Credit Card Fraud Detection*. [online] Kaggle.com. Available at: <https://www.kaggle.com/mlg-ulb/creditcardfraud> [Accessed 13 August 2020].

8Preda, G., 2018. *Credit Card Fraud Detection Predictive Models*. [online] Kaggle.com. Available at: <https://www.kaggle.com/gpreda/credit-card-fraud-detection-predictive-models#Data-exploration> [Accessed 13 June 2020].

9Charminda D, 2020. *Credit Card Fraud Detection: A Detailed Study*. [online] Kaggle.com. Available at: <https://www.kaggle.com/chandrimad31/credit-card-fraud-detection-a-detailed-study> [Accessed 13 July 2020].

10Escapa, C., 2018. *AWS Analytics Services Explained: From Data Lakes To Machine Learning | Amazon Web Services*. [online] Amazon Web Services. Available at: <https://aws.amazon.com/blogs/apn/aws-analytics-services-explained-from-data-lakes-to-machine-learning/> [Accessed 7 August 2020].

11Sharma, A., 2020. *How To Use Google Colab For Deep Learning And Machine Learning*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/03/google-colab-machine-learning-deep-learning/#1> [Accessed 7 August 2020].

12Microsoft Tutorial, 2020. *ML Studio (Classic): Quickstart: Create A Data Science Experiment - Azure*. [online] Docs.microsoft.com. Available at: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment> [Accessed 7 August 2020].

13Kipp, A., 2017. *Shiny - Shinyapps.Io - Getting Started*. [online] Shiny.rstudio.com. Available at: <https://shiny.rstudio.com/articles/shinyapps.html> [Accessed 7 July 2020].

14Highsmith, J., 2007. *Agile Project Management*. 23rd ed. Boston, Mass.;Munich[u.a.]: Addison-Wesley.

15Beck, et al., 2001. *Manifesto For Agile Software Development*. [online] Agilemanifesto.org. Available at: <https://agilemanifesto.org/> [Accessed 7 August 2018].

16Agile Alliance, n.d. *What Are User Stories?*. [online] Agile Alliance. Available at: <https://www.agilealliance.org/glossary/user-stories> [Accessed 7 March 2018].

17Sachdeva, S., 2016. Scrum Methodology. *International Journal Of Engineering And Computer Science*,.

18Huether, D., 2017. *The Definition Of Done*. [online] LeadingAgile. Available at: <https://www.leadingagile.com/2017/02/definition-of-done/#:~:text=The%20definition%20of%20done%20(DoD,%2C%20team%2C%20or%20consuming%20system.> [Accessed 7 March 2018].

19Verner, T., 2019. *Microsoft Azure AI Engineer: Developing ML Pipelines In Microsoft Azure*. [online] Pluralsight.com. Available at: <https://www.pluralsight.com/courses/microsoft-azure-developing-ml-pipelines> [Accessed 7 June 2020].

20Deliwala, S., 2019. *Streamlit 101: An In-Depth Introduction*. [online] Medium. Available at: <https://towardsdatascience.com/streamlit-101-an-in-depth-introduction-fc8aad9492f2> [Accessed 7 August 2020].