

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](#_Toc48931159)

[1. Introduction 4](#_Toc48931160)

[1.1. What the Project Aimed to Deliver 4](#_Toc48931161)

[1.2. How the Project Delivery was Implemented 4](#_Toc48931162)

[2. Background / Literature Review 5](#_Toc48931163)

[2.1. Credit Card Fraud Detection: Further Research on Predictive Models 5](#_Toc48931164)

[3. Requirements: Specification and Design 6](#_Toc48931165)

[3.1. High Level Project Requirements 6](#_Toc48931166)

[3.2. Project Architecture Diagram 6](#_Toc48931167)

[3.3. High Level Project Design 7](#_Toc48931168)

[3.3.1. Prototype Development 7](#_Toc48931173)

[3.3.2. Final Project Deliverable 7](#_Toc48931174)

[4. Project Implementation(1) – Azure Modelling 8](#_Toc48931179)

[4.1. The Machine Learning Workflow 8](#_Toc48931180)

[4.2. Credit Card Fraud Dataset – Analysis and Preparation 8](#_Toc48931181)

[4.3. Credit Card Fraud – Building the Azure Model 8](#_Toc48931182)

[4.4. Credit Card Fraud – Deploying the Azure Model 8](#_Toc48931183)

[5. Project Implementation(2) – Shiny R Dashboard UI 9](#_Toc48931184)

[5.1. Data Visualisations in a Shiny Dashboard 9](#_Toc48931185)

[5.2. Credit Card Fraud – UI to Check Fraud Predictions 9](#_Toc48931186)

[5.3. Shiny UI – Hosted Application 9](#_Toc48931187)

[6. Testing and Results 10](#_Toc48931188)

[6.1. User Story ‘Demos’ – Test Results and ‘Feedback’ 10](#_Toc48931189)

[7.1.1. User Story 4: Initial Data Modelling – Review and Evaluation 10](#_Toc48931194)

[7.1.2. User Story 5: Basic Shiny App – Review and Evaluation 10](#_Toc48931195)

[7.1.3. User Story 6: Integrated Prototype – Review and Evaluation 10](#_Toc48931196)

[7.1.4. User Story 7: Enhanced Modelling – Review and Evaluation 10](#_Toc48931197)

[7.1.5. User Story 8: Enhanced UI – Review and Evaluation 10](#_Toc48931198)

[7.1.6. User Story 9: Presentation Preparation – Review and Evaluation 11](#_Toc48931199)

[6.2. Final Project Assessment 12](#_Toc48931200)

[6.3. Project Plan 2020: Final Status – 25th September 2020 13](#_Toc48931201)

[7. Project Location and User Guide 14](#_Toc48931202)

[7.1. Credit Card Fraud Application: Prototype Location 14](#_Toc48931203)

[7.2. Credit Card Fraud Application: User Guide (Final Project) 14](#_Toc48931204)

[8. Project Conclusions 15](#_Toc48931205)

[8.1. Where Project Goals Achieved? 15](#_Toc48931206)

[8.2. Future Design/Deployment Considerations 15](#_Toc48931207)

[9. Appendices 16](#_Toc48931208)

[9.1. Azure Generated Code Segments 16](#_Toc48931209)

[9.2. Shiny R Application Code Files 18](#_Toc48931210)

[13.2.1. Diagram: The RStudio Cloud Environment 18](#_Toc48931219)

[13.2.2. The Shiny UI Code 18](#_Toc48931220)

[13.2.3. The R Code Parsing Data and Invoking UI 18](#_Toc48931221)

[9.3. Azure Machine Learning Classic Studio Experiments 19](#_Toc48931222)

[9.4. Credit Card Fraud Datasets 20](#_Toc48931223)

[10. References / Bibliography - Interim Report 21](#_Toc48931224)

# 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

The Interim Report…

<image>

One-pager..

As of mid-August 2020, ..

* The...

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

Describe my dataset in the context of the research….

Just a few bullet points..

# 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

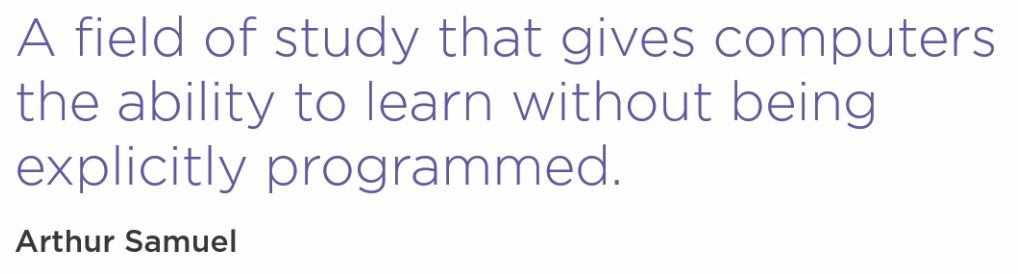
A significant amount of training and reference material I had access for guidance on using the Azure Machine Learning Studio came from Pluralsight courses.

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.

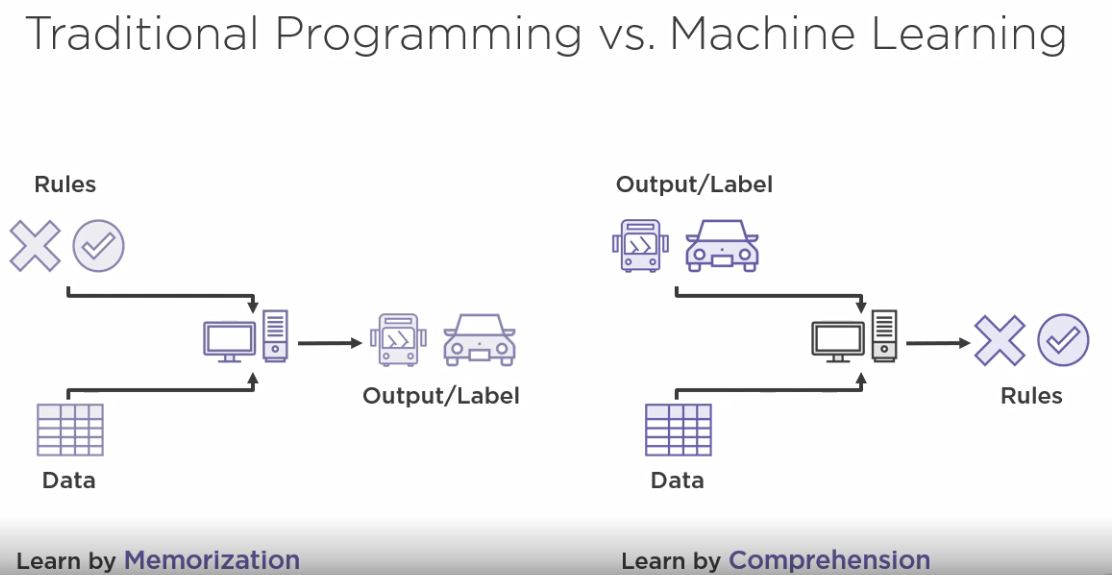
To start with a quote…’what is Machine Learning?’

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



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

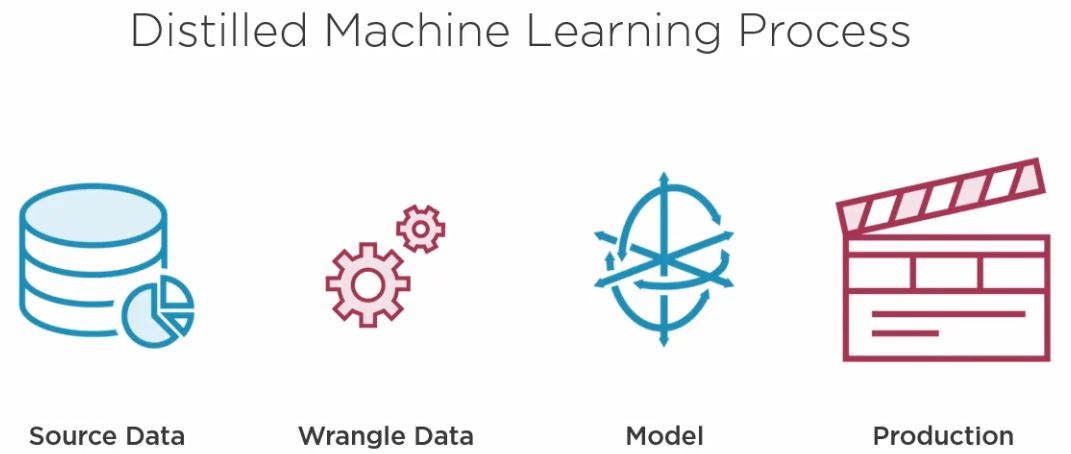
Figure: Traditional Programming v Machine Learning - *Reproduced Image from Pluralsight* (Rhodes, 2020)



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.

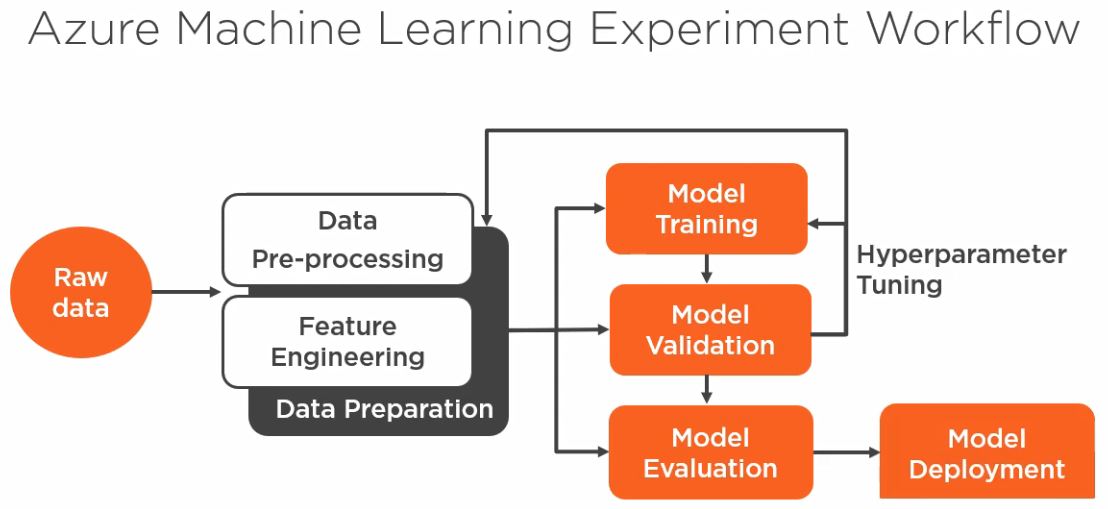
This is a simplified diagram of how the Machine Learning process is applied.

*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.

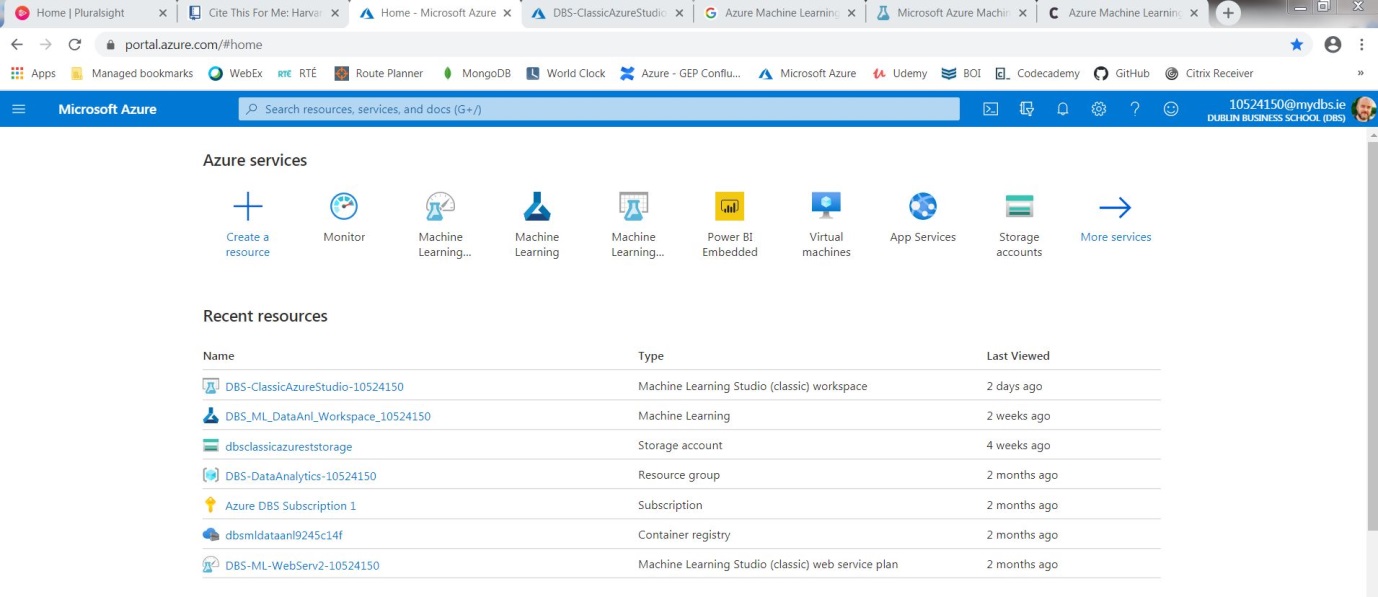
In brief; a description of Azure Workspaces can be found here; <https://docs.microsoft.com/en-us/azure/machine-learning/concept-workspace>

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>

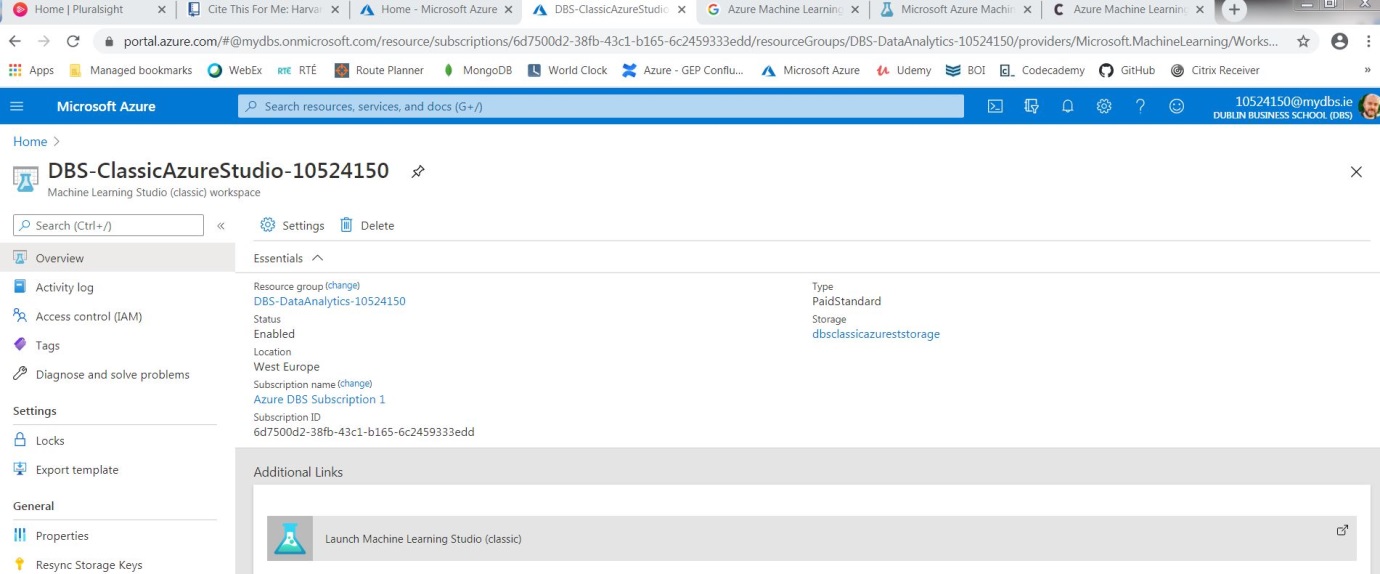
To access the Azure Machine Learning Studio (Classic) where I developed my project the first step is to log onto the Azure Portal, which I set up with my DBS account.

*Figure: Azure Portal (my DBS account)*



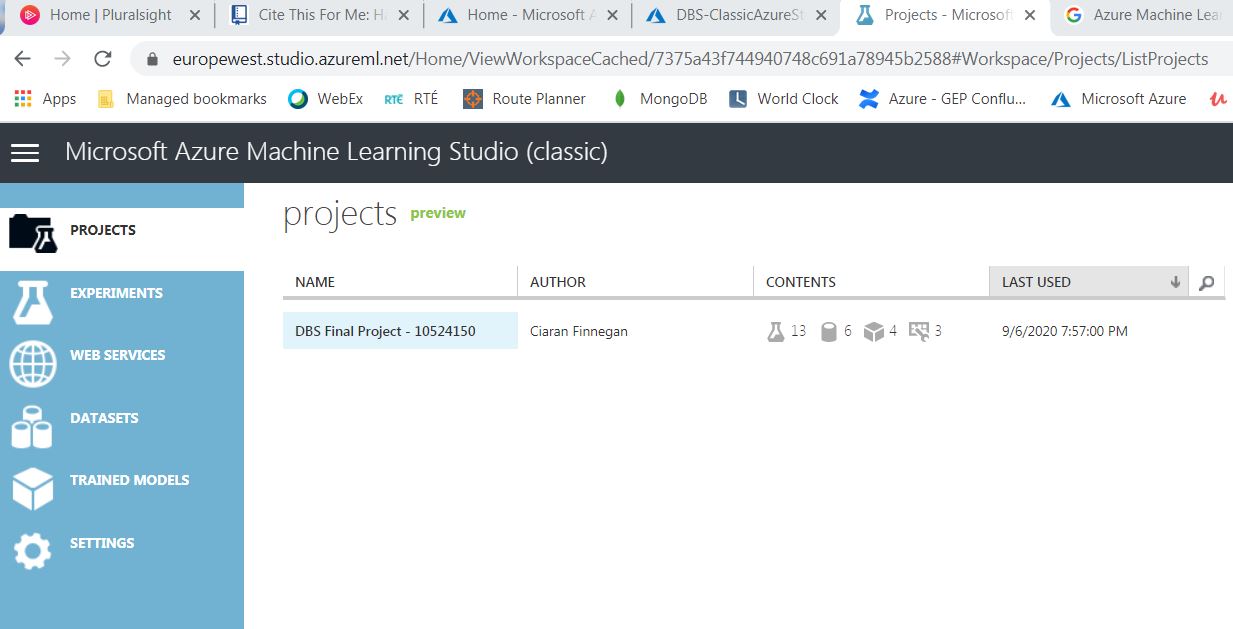
A workspace has been created by me for a Machine Learning Studio (Classic) service.

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.

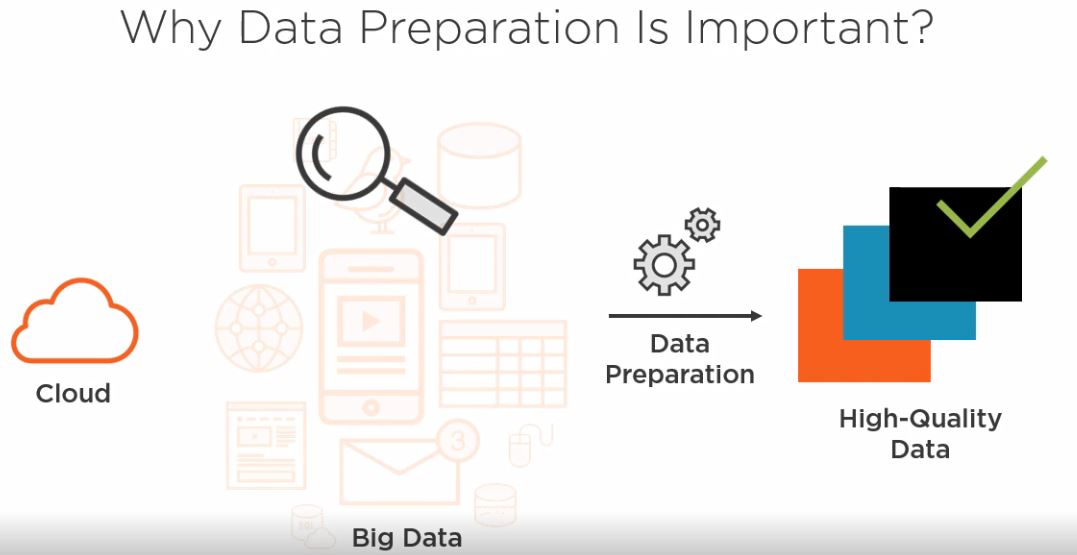
*Figure: Overview of Azure ML Studio (classic) environment for this project – Experiment View*

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 number 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 how…

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:..

This..

### Experiment 5:..

This..

### Experiment 6:..

This..

### Experiment 7:..

This..

## Credit Card Fraud – Deploying the Azure Model

The..

<images>

# 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’

The prototype for this project..



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

Goal:

Assessment of robustness of code and functionality delivered:

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

Goal:

Assessment of robustness of code and functionality delivered:

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

Goal:

Assessment of robustness of code and functionality delivered:

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

Goal:

Assessment of robustness of code and functionality delivered:

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

Goal:

Assessment of robustness of code and functionality delivered:

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

Goal:

Assessment of robustness of code and functionality delivered:

## 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*

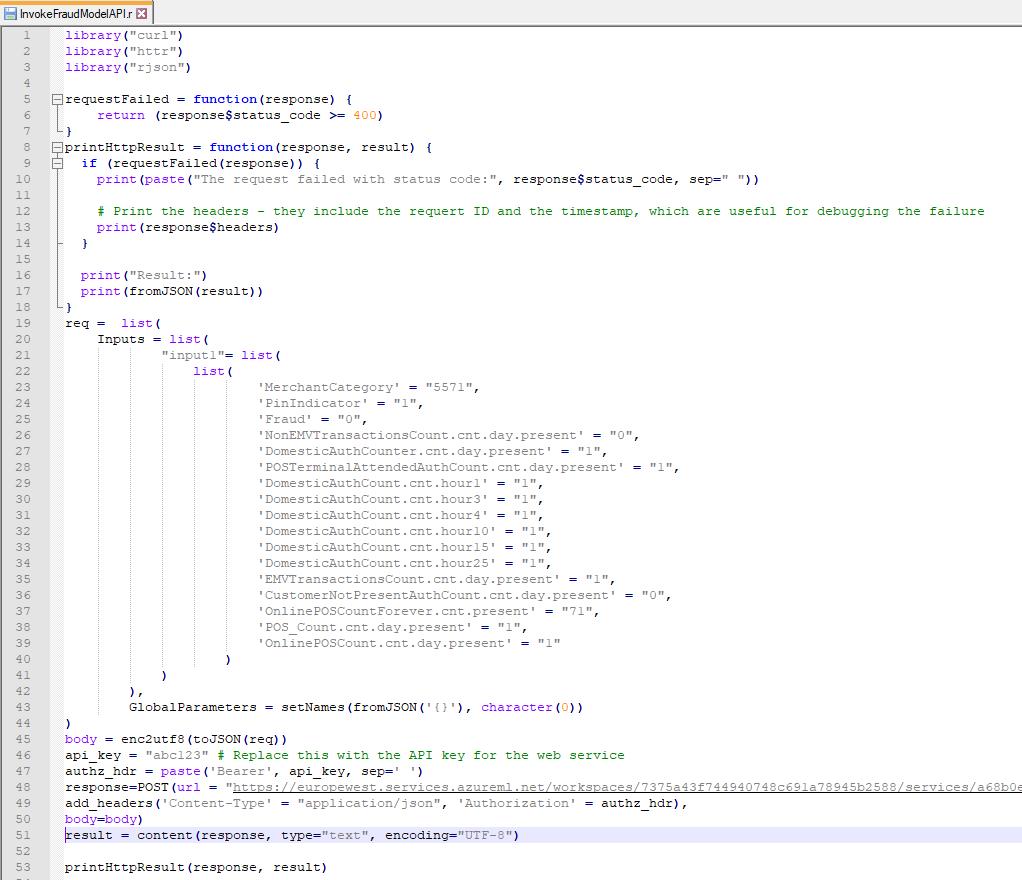
A screenshot of a social media post

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

## 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].