Data Analyst Portfolio

[About: 1](#_Toc47655440)

[Project 1 – Customer Fraudulent Activity Time Series Analysis – Financial Crime Analytics 1](#_Toc47655441)

[Business Problem 1](#_Toc47655442)

[Workflow 2](#_Toc47655443)

[My Role 2](#_Toc47655444)

[Data Sources 2](#_Toc47655445)

[Competencies I Have Met 2](#_Toc47655446)

[Outcome 3](#_Toc47655447)

[Project 2 – Internal Package to Automate Impala and Hive queries within a Cloudera Environment. 3](#_Toc47655448)

[Business Problem 3](#_Toc47655449)

[Workflow 3](#_Toc47655450)

[My Role 3](#_Toc47655451)

[Data Architecture 3](#_Toc47655452)

[Outcome 4](#_Toc47655453)

[Reflective Journal 1 – Mortgage Churn Model – Stakeholder Management 4](#_Toc47655454)

[Reflective Journal 2 – Data Governance 5](#_Toc47655455)

[Reflective Journal 3 – Data Hackathon 5](#_Toc47655456)

# About:

My name is Shaan and I am currently a Data Science Degree Apprentice at Santander UK, I have had rotations within the Core Data Science Team of the bank as well as within Financial Crime. I have a strong interest in Data Science, specifically Reinforcement learning as well as Financial Markets. After this apprenticeship.

# Project 1 – Customer Fraudulent Activity Time Series Analysis – Financial Crime Analytics

## Business Problem

In Santander, as one of the major banks within Europe and the second largest bank in Spain, we often have customers who gain money via illicit transactions and attempt to ‘hide’ their trail via ‘money laundering.’ We are attempting to minimize the number of fraudulent activities and to detect which months and which specific customers are involved in money laundering, to uphold the reputation of the bank.

## Workflow

I was working with transactional data for a select group of ‘suspicious’ customers who were identified, and we wanted to see whether spending habits had significantly changed in terms of our customers withdrawing and depositing money. This would be used to identify which months we would specifically look at. Then from the specific months, we would continue our analysis to look at which specific customers were likely to be involved in fraudulent activity.

## My Role

Within the project I was tasked with:

* Liaising with various internal/external stakeholders, gathering the project requirements as well as presenting my analysis to the respective stakeholder of the project.
* Performing analysis, to see which months had the most unusual spending activities out of the group of customers we had initially identified.
* Creating an ARIMA model to predict the results for future months.
* Presenting the results to the respective stakeholders, as well as informing my team of what months fraudulent activity was most prevalent within the identified group.

## Data Sources

Data Lake

Cloudera

Using my package, I extracted the data from the data lake into Cloudera, before importing the data into a Jupyter Notebook to perform the analysis on my local machine. I had to uphold data governance and the

## Competencies I Have Met

* Identify the input data necessary based on requirements elicitation for: Internal Data
* Collect data from a data source
* Apply and refer to the organisational policies relating to your use of data
* Apply different data mining tools (Excel, Tableau, SQL, R, Python etc) to large datasets to find patterns
* Apply methods for cleaning incorrect, inconsistent, redundant and incomplete data
* Apply time series analysis to your dataset to determine trends
* Apply time series forecasting to a dataset to predict a future value
* Undertake an analysis which produce actionable insights
* Explain the quantitative insights in a piece of analysis
* Present the outcomes of an analysis to a stakeholder
* Summarise the outcomes of an analysis using different communication methods
* Justify the use (or lack of) Big Data technologies in your analysis
* Describe the data architecture that supports your analysis
* Comment on the relevance of a company policy/standard/model/rule in your analysis
* Apply data cleansing principles to your datasets Structured

Evidence of the above can be found in the Jupyter Notebook or within the document.

## Outcome

As a result of the project, we have been able to narrow down our analysis. When I presented the findings to the external stakeholders of the project, they agreed that it would be necessary to narrow down our search from January 2019 – April 2019, due to the sudden change in the forecast. As a result, the rest of my team are currently working on identifying which customers can be identified within the three-month period that I identified where transactions suddenly changed.

# Project 2 – Internal Package to Automate Impala and Hive queries within a Cloudera Environment.

## Business Problem

Within the bank there are multiple functions which need to extract data from the data lake. Including but not limited to Business Analysts, Quantitative Analysts and Data Scientists. However, I found that when ingesting data into the data lake and retrieving data from the data lake, it was a fairly lengthy process and unattainable if you didn’t have prerequisite knowledge of the Data Architecture as well as Hadoop. Hence, I decided to make an easy to use package.

## Workflow

The project wasn’t initially assigned to me, more of a task that I identified as an area of improvement for the bank, and I knew that by automating the task of running queries within Cloudera using the Hadoop HDFS, I could massively increase efficiency across the bank. I decided to initially improve my knowledge of Hadoop and the data architecture used within Santander. Then I tried applying my newfound knowledge to projects which required me to interact with the data lake. Then, I adjusted the parameters for the package, depending on which parameters were most likely to be used by others in the bank. After developing the package, I presented it internally within my team, before writing up the final documentation and releasing the package across the bank.

## My Role

As the sole creator of the package, my responsibilities included:

* Initial testing of the package and the sole developer.
* I was required to understand the data architecture as well as the needs a goal of Santander to compete with FinTech companies who don’t have the problem of legacy systems.
* Justifying with internal stakeholders and my team, for the package to be released across the bank.
* Optimizing the speed to run the package within CDSW
* Constantly maintaining the package and taking into consideration other developers opinions on how the package could be improved.

## Data Architecture

DC1

DC2

The package allows users, to specify the database, hence not restricting any employees within the bank to one cluster within the hadoop ecosystem.

## Outcome

As a result of the package, the Data Science community within the bank (150 people) can access and use the package, automating the manual process of obtaining and ingesting data into the data lake. As well as allowing other developers without the prerequisite knowledge of Hadoop to use the data lake without a UI. The overall efficiency increase was around 40% compared to the previous method of running the queries

# Reflective Journal 1 – Mortgage Churn Model – Stakeholder Management

Within my two rotations so far, within the core data science team of the bank and in financial crime analytics, I have been exposed to various internal and external stakeholders.

Within the mortgage churn team, I was encouraged to gather project requirements from internal stakeholders via email and face-to-face meetings. Theses meetings would be to establish the business context of the benefits of the project. From these meetings we gathered that our aim was to reduce the number of customers taking their mortgage from Santander to another bank by 20% when they came to the end of their fixed rate period and enter variable rate on the respective loan.

Once we had gathered the requirements from the internal stakeholders, it was time to program our first draft of the churn model, I was primarily involved in the development of the mortgage churn dashboard, whereas my line manager was more involved with the development of the model using Machine Learning. By collaborating with my manager, we prioritised the dashboard, due to our external stakeholders requiring an overview of our project which would be presented via the dashboard.

We were also required to extract data from various tables and create a single table in SQL. This was quite a lengthy process as our first iteration we input the wrong cut\_off\_date for one of our tables. Hence, we were predicting for the last four months rather than the last three months, which required us to adapt our time schedule and focus on the creation of the tables, which would be input into the dashboard. We were able to meet the deadline for our first presentation to external stakeholders regarding the churn dashboard, I also presented the dashboard to members of my team, to give an overview of the project. I gave my recommendation, in that we should contact the individuals who were identified within the model as ’high risk’ of leaving Santander and offer to extend their fixed rate an additional two years.

I was required to choose different methods in presenting to both my team and the external stakeholders, with my team I chose to present my code and the dashboard. Whereas with the external stakeholders, I chose to present using a presentation and the dashboard, due to my team comprising of Data Scientists who understood the code, whereas the external stakeholders were more interested in the commercial value of the project.

# Reflective Journal 2 – Data Governance

Within my journey as a Data Science Degree Apprentice at Santander, I have been introduced thoroughly into the data protection legislation as a bank and to meet our customers needs, how it is a necessity for us to abide to GDPR.

One of the primary reasons, is that as banks, failing to abide to regulations regarding data protection, could put our customers/clients at risk and are at risk of a potential sanction. All employees working with sensitive data are required to go through data protection training, which has made me aware of the consequences for the bank.

I first asked my new team any regulations regarding data protection, that I should be aware of, knowing that I work with sensitive client data, that if I were to breach confidentiality it could be disastrous to the bank.

I also know that at Santander, despite our aim to compete with the emerging FinTech companies, we still need to comply with regulations such as GDPR, whilst ensuring the service we provide to our customers (retail bank) or companies (CCB/CIB) is the very best. The service we provide is enhanced by innovation, whilst Data Science is a key aspect of the innovation, we cannot afford to lose our clients trust in our service by breaching GDPR. By exceeding our customers expectations and ensuring that GDPR is complied with, no confidential data is lost nor used for purposes other than

I actively have tried to attend various data hackathons and network with other Data Scientists within the FinTech space and have asked numerous individuals how we can use data effectively commercially within a bank, whilst complying with regulation. One notable person I have asked was the director of data science at KPMG, who gave me a great introduction into how innovation and regulations are almost a scale, which balance one way, affect the other way, and the aim is to in fact balance innovation and regulation as much as possible.

# Reflective Journal 3 – Data Hackathon

The data hackathon ran internally within Santander was one of my first events that I attended within the Data Science community at Santander. The aim of the day was to get those interested in Data Science regardless of there role an insight into how data can be leveraged by the bank as well as offering technical training to colleagues who felt it may benefit their role.

In terms of technical training, the session comprised of various activities from theory to practical and I gained an insight into the different types of industry Machine Learning algorithms. I can confidently say that in terms of unstructured data, there are various formats such as video, social media or text data.