# INNOVATION STRATEGY CLOSE-OUT REPORT

| | |
|---|---|
| PROJECT TITLE | Data Analytics for Revenue Protection Fraud Detection |
| PROJECT OWNER | Mick Finn, Data Analytics Manager, ESB Networks |
| CONTRIBUTOR(S) | Seamus Gray |
| INTERNAL DOCUMENT NO | DOC-161019-FET |
| VERSION | 1.1 |
| DATE | 6th June 2019 |

## BRIEF OVERVIEW OF PROJECT & EXPECTED BENEFITS

It is estimated that €20-€30 million of ESB Networks' revenue is lost each year through customers tampering with electricity meters and cables. The revenue protection team based in Tralee engaged with the ESB Networks Data Analytics team to determine how they might be able to assist them in using their data more effectively.

This project team came up with 2 main objectives:

A) To use Data Analytics to predict fraudulent behaviour based purely on an analysis of historical data. Using advanced data science techniques, we hoped to proactively identify suspicious patterns that will allow Revenue Protection team to proactively investigate.
B) Develop a more integrated reporting solution (by local area and transformer sub-station including graphs) for the Revenue Protection team to allow them to do more detailed analysis themselves. The existing reporting solution at the time could not do this analysis by local area.

There were two main benefits expected:

1. Reduce the amount of lost revenue due to theft.
2. Reduce the amount of unsafe practices that members of the public engage in while tampering with meters.

## RESULTS

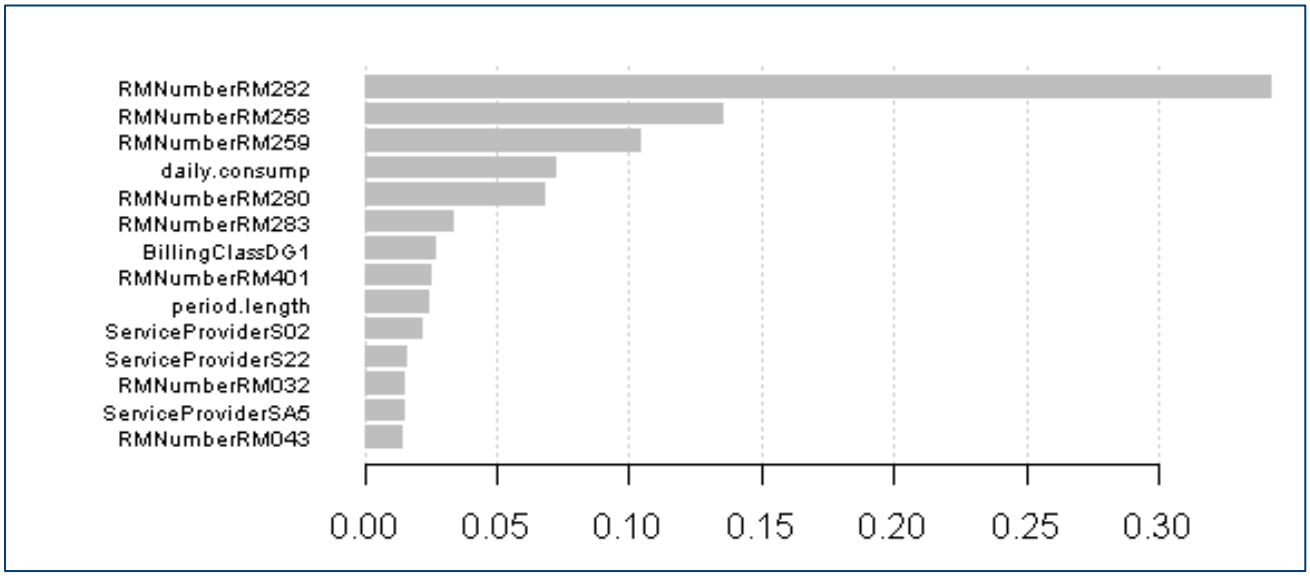
A pre-requisite to both objectives was to gather all the required data into one central system.

Working closely with the Revenue Protection team, the Data Analytics team extracted the source data into a central data mart (SQL Server) for analysis to identify key trends important to the team. This proved to be a difficult task due to the dispersed nature and volume of the data.

The first deliverable was a statistical model developed to try and proactively highlight potential areas of fraud based on analysis of historical data. This proved a very challenging piece of work.

Initial attempts to look at the problem purely based on consumption history proved inaccurate. A number of factors like estimate readings and genuinely erratic consumption patterns (e.g. student houses rented out during the year but maybe have nobody living in them during the summer and holiday periods) skewed results. This unsupervised learning model was deemed not reliable enough by the business. The majority of the identified potential fraudulent sites were false positives.

Based on further reading of fraud studies done in other utilities (and indeed other industries like credit card fraud) a more complicated supervised learning model was developed (extreme gradient boosting). This model considered a lot more factors and attributes e.g. meter type, seasonality etc. The model was trained based on these attributes for cases where known fraud had previously been identified.



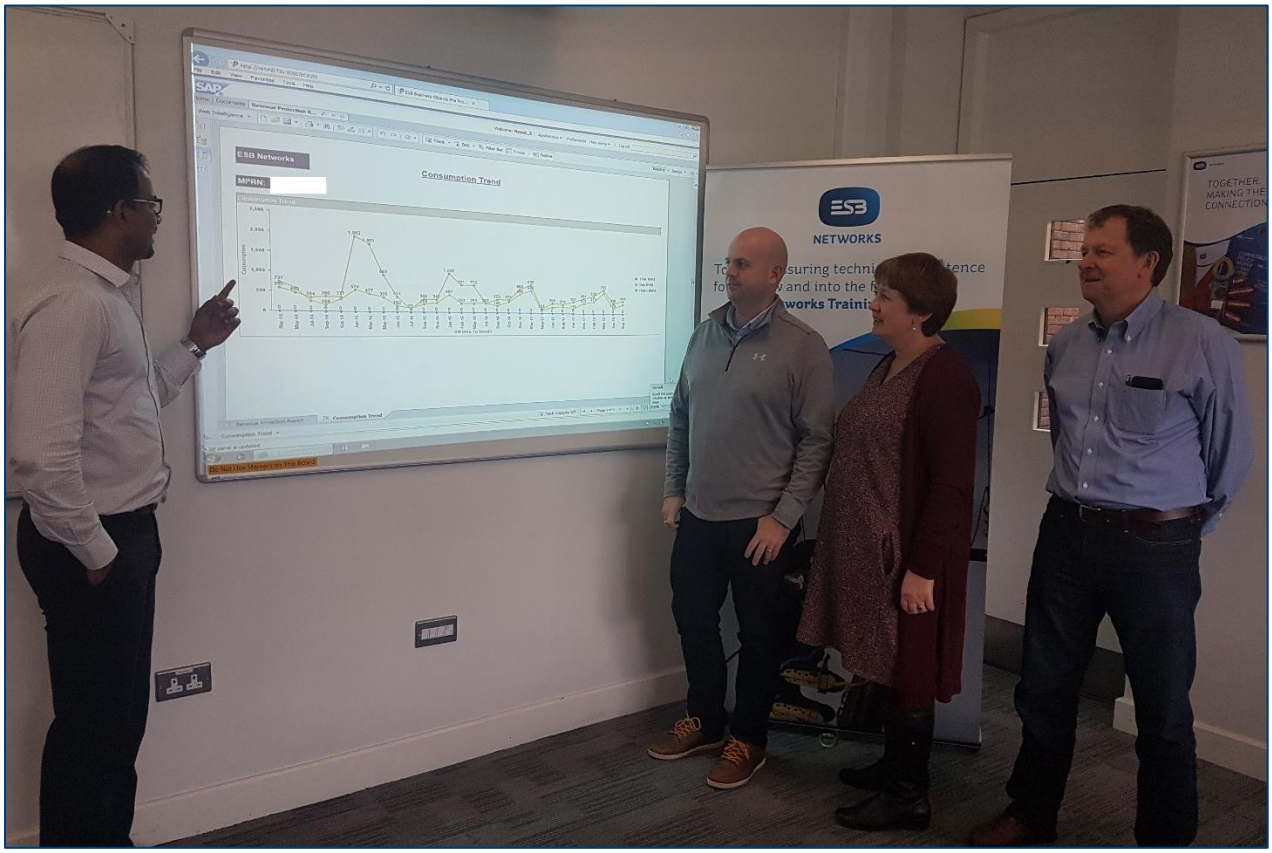

*The graph above shows the attributes deemed most relevant and their statistical significance*

The model was developed and executed for domestic customers in a predefined small geographical region only.

The initial accuracy of the model on the training data set was 92% for the true positives. This means that when we run the model on historical data for a time period where we have the actual results, it successfully predicted 92% of the cases (this means it did NOT predict 8% of the cases). As is standard practice for supervised machine learning, the model was trained on ¾ of the dataset (150k records) and tested on the rest of the data set (50k records).

The full list of premises for the region was subsequently run through the model. From this run a list of 329 possibly suspicious accounts were identified.

For the second deliverable, a new report was developed which for the first-time enabled Revenue Protection to produce consumption data by local area and Transformer Sub Station number (including trend graphs). This reporting solution allowing analysis by local area was successfully delivered and it has removed a lot of manual data manipulation work from the team and allowed them to spend more time on high value activities like the actual fraud analysis.




*Pictured Left-to-Right – Shivananda Nayak (BI & Analytics team), Keith McCarthy, Martina Hickey, Seamus Gray (Networks Revenue Protection team)*

In the photo above we can see Shivananda explaining to the Revenue Protection team the work that the project completed in order to help view their data through a different lens.

## LEARNINGS

1. This is a non-trivial problem which has a lot of complexity from simply gathering all the relevant data to coming up with a model which is accurate enough to usefully use
2. It was not possible to complete this model for quarterly hour (QH) customers. This is because when we find fraud on QH meters, we overwrite the readings with a more accurate readings of what should have been billed. This is to try and recoup the money lost from this theft of revenue. However, it means that we no longer have the historical data on what suspicious consumption patterns look like for QH customers
3. This is a data intensive exercise especially if we want to execute for all accounts in all of the country in a repeatable fashion. Effort would be to automate the whole data gathering and update process if / when the model is integrated fully into the Revenue Protections business processes
4. There is a willingness within the business to try new methods and techniques in order to try and be more effective and proactive in their day-to-day job
5. Statistically, prepayment meters have shown to have a higher likelihood for fraud. However, a revised anti-tamper (magnetically at least) meter was introduced in the last year or two. This has the effect of changing the patterns of tampering and therefore the model needs to be amended to take account of this. [This is a common problem with machine learning algorithms, as behaviours or other factors change the model likewise must change in line with this.]

## BENEFITS REALISED/VALIDATED

1. The delivery of an improved analytical report to the Revenue Protection team means that they can themselves now do more in-depth and targeted analysis. This saves them time as up to now they would have had to spend a lot more time doing manual data manipulation to do the analysis they required to do their day job. It also allows them to more easily run 'campaigns' in their search for fraudulent behaviour
2. The revenue protection team now also have a greater understanding of what attributes are relevant in assessing likelihood of fraud (even if they are manually doing it)
3. A model can be produced to support any future revenue protection campaigns that will narrow down the search for suspected fraudulent behaviours significantly
4. Decrease the amount of fraud happening, estimated at €20-30 million per annum.
5. Reduce the amount of unsafe and dangerous behaviours being carried out by the public during attempted fraud cases

## NEXT STEPS – BAU, TRANSFER OF OWNERSHIP

1. The model output is still being investigated by the Revenue Protection team (329 cases is a large number to be investigated). While some basic 'desk checks' can be done to see if there is a genuine reason for the unusual patterns the final checking requires a physical visit to the premise by a Network Technician. This is a labour intensive and costly exercise.
2. Proposal on how to 'productionise' (to allow further rollout) this statistical model has been developed and is awaiting the outcome of the testing to prove that the business cases exist for further rollout.
3. Advanced reporting solution (Planner Grp/Sub No) is already in production and standard support model is in place between Networks business and IT support teams

## FINAL TIMELINES (REASONS FOR ANY DELAYS IF THEY OCCURRED)

1. Advanced reporting solution went live in early September 2018 as planned
2. Model development / evaluation was late by approx. 1 month due to the complex nature of the analysis and the change in approach after first cut model results were deemed not accurate enough
3. Testing to complete Q4 2019 and productionising Go / No Go decision to happen then

## FINAL COSTS

Final costs were €150,000 – all time costs for developers and business input.

No software or hardware costs were incurred. Either open source or existing technologies were used.

Depending on productionising decision some additional costs may be incurred.