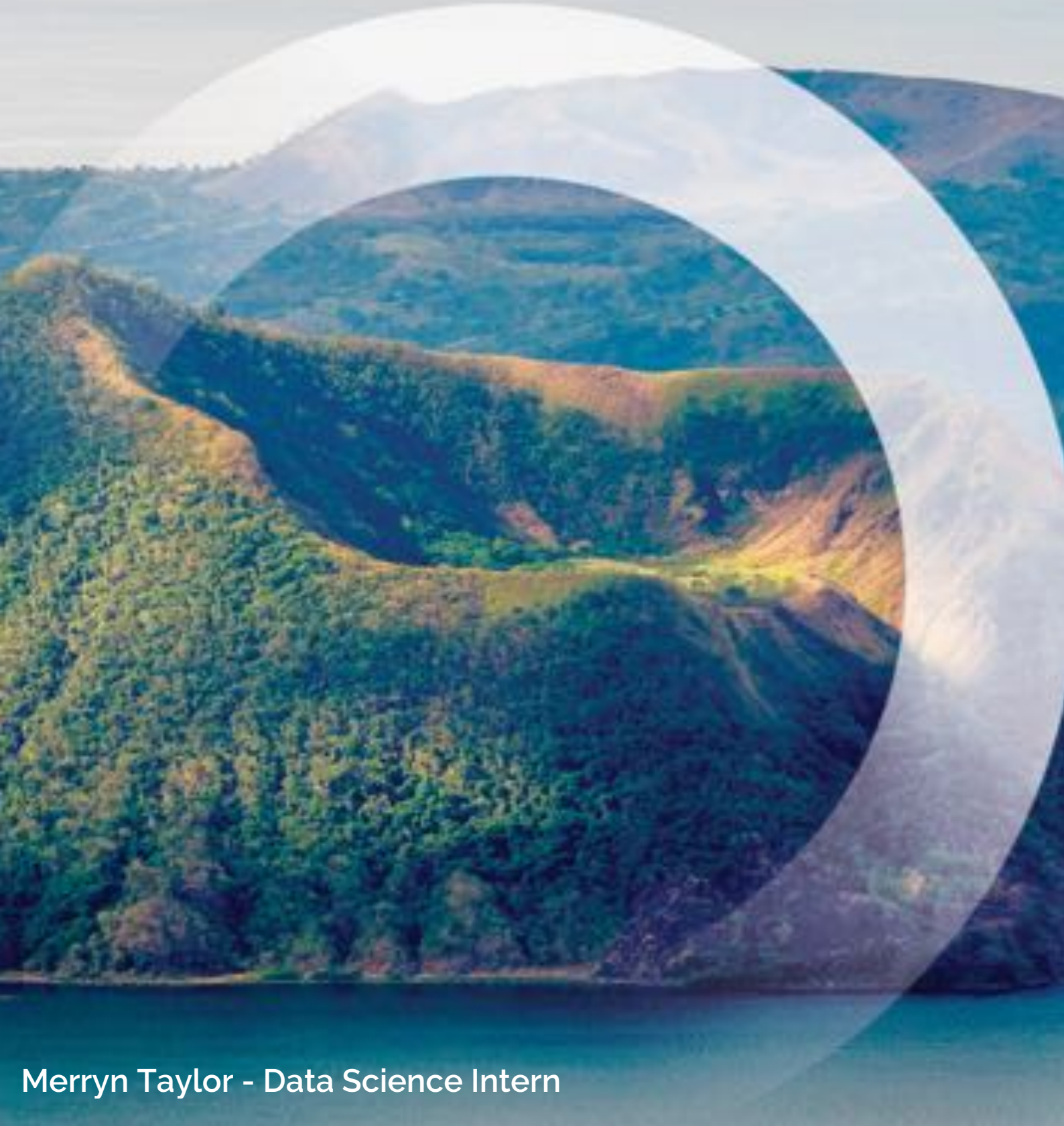
LOSS MODELLING FRAMEWORK

ANALYCAT
AI.MATH.SUPERCOMPUTING

Bot SUE data reconstruction test



Merryn Taylor - Data Science Intern

A BOT NAMED SUE

In order to investigate the potential of SUE for data augmentation in cat modelling, a set of exposure data (address and construction details of more than 12 thousand buildings in Melton Mowbray provided by the Ordnance Survey) was taken and the field containing building class was removed. SUE was then used to attempt to reconstruct that data from some of the remaining information. The steps taken are shown and outlined below.

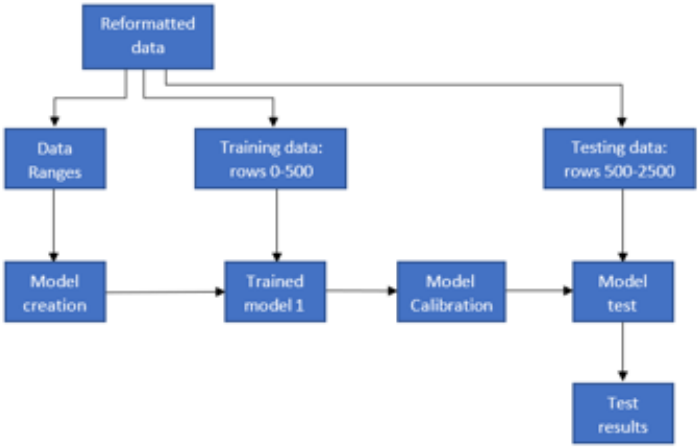


Figure 1 - shows the data flow of the testing method for SUE

Data Processing

In order for SUE to be able to process the data it first had to be processed into the right format. This involved removing data fields where there were only unique values (such as ID columns) or columns that had very few data points in them (such as business names columns). Data fields were also removed that would make manual entry of the variables into SUE unfeasible. For example, the postcode field could have potentially been useful however, with over 300 categories to manually enter this would have taken too long.

Finally, a copy of this new data set was created, and ranges and unique categories were calculated.

Model Creation

With the data processed, a SUE model can now be created. This involved manually entering all of the categories for building class into the output variable, and manually entering all of the categories and ranges into new input variables for each of the remaining data fields.

Model Training

The model can now be trained. In order to train the model a sample of 500 rows of data was taken from the processed data. 500 was chosen in order to give SUE enough data about the rarer types of building. With less samples, it became hard for SUE to identify buildings such as churches and generic commercial properties. These 500 samples were uploaded to SUE using a generated import template which was then populated from the processed data file.

Model Calibration

The final step before SUE can be used is to calibrate the model. This involves changing the number of samples SUE uses for each decision as well as changing the weights of each of those samples. In this case SUE was configured to use 3 samples with a smooth weight fall off. Changing the levers in this case did not provide any performance benefit. The final mean forecasting error before testing was 7.8%.

Model Usage

With the model trained it can now be tested. In order to get a good idea of SUE's predictive ability a sample of 2000 rows of data were taken from the processed data set. This large quantity was taken to investigate SUE's ability to recognize rarer building classes. The data was imported into a generated import template and then uploaded. [1] Once this was done, SUE could be queried. When queried over 2000 samples, SUE's was completely correct in over 91% of samples on 500 samples. Upon reviewing the data, it was also clear that a lot of the misclassifications made by SUE were within the same general category i.e. small errors similar to the correct answer. This could also possibly be due to inconsistent recording in the original data or that SUE requires more samples of those types of building in order to identify them more reliably.

Conclusions

An important point to consider is that of the 20 possible types, the proportion of the properties that were of the RD (generic residential) class. There were only 12.9% of properties that did not fit within this class. This is an overwhelming majority and therefore might skew the results.

Another important point to consider is the data processing that was done in the beginning. This data processing was not done by an expert and therefore important or useful data could have been removed. Moreover, some of the data fields that were not modelled due to their difficulty to manually enter could have also improved SUE's performance.

[1] Importantly though, before SUE could be asked to calculate building classes for all of these samples, around 45 points of data had to be manually edited due to a bug with SUE's internal representation of large numbers.

Therefore, if this test was to be done again it should be done with a more complete data set in order to get an even better idea of SUE's capabilities.
Overall though, this test demonstrates the ability of SUE to reconstruct missing data in datasets reasonably accurately from very little data. In the future an automated approach to data field removal would drastically improve the usability of the system. Furthermore, using SUE's visualisation it was clear that some of the fields used could also be removed, due to their low impact on the outcome. This highlights SUE's ability to reduce the dimensionality of data.
The reader must note that this test was executed with limited time and resources and is purely indicative. A more comprehensive test would be needed to ensure a complete review, but for this single test, the numbers are as above.

ABOUT ANALYCAT

Analycat has its roots in the complex world of modeling natural disasters for the financial world. Since 2008 their various software products have been used to prepare banks, insurers and countries for the worst possible financial blows from nature. Today they have taken that sophisticated knowledge of mathematics and programming and applied it to artificial intelligence.

Analycat sets out to produce the type of software that the market demands but doesn't get – fully transparent platforms that provide default, state of the art quality actuarial and scientific data, affordability, usability, strong graphics and flexibility for in-house use of informed and risk-minded stakeholders.

Find our more http://analycat.com
Contact info@analycat.com