UNIVERSITY OF CAMBRIDGE
Department of Engineering

CONSTRUCTION ENGINEERING MASTERS DISSERTATION ABSTRACT

Project Outcome Prediction Using Machine Learning

Engineering consultancies have long fought to maintain sustainable profit margins. The bespoke, complex nature of their work, performed in an often-contentious and cash-strapped industry, makes the task harder. Significant effort is expended to timeously identify engineering design projects that will under-perform. In parallel, digitisation of the engineering industry has failed to bring the expected benefits but has resulted in the vast majority of the industry's records being digitised.

We take inspiration from the value that pioneering companies in other industries have extracted from their data. The broad field of artificial intelligence and specifically the application of machine learning (ML) have proven the ability to extract patterns from vast troves of data, patterns which tell the story of those companies' customers, products, projects and more. The value provided to these companies by ML ranges from improved awareness of their business, to philanthropic spin-off operations, to whole areas of new revenue generation. Could engineering consultancies do the same? The engineering industry's digital transformation, whilst unsuccessful by some measures, has resulted in an abundant data about projects, customers and people. The need to timeously identify problem projects is more pressing than ever. Should it not be possible to leverage the phenomenal predictive powers of ML to assist with identifying these projects?

This study had two components: review the myriad data stores held by a large consultancy to understand their usability, then to attempt to make predictions about engineering design project outcomes using that data. The initial phase was undertaken by a series of investigations into obtaining data. The second phase comprised taking timesheet data from 7,996 engineering design projects and employing ML techniques to predict how those projects' profit margins varied over their course.

The first phase identified that whilst there is a huge amount of data available, a relatively small proportion of is structured. More positively, the most useful data is already in structured form and even unstructured data is described by metadata, which is itself structured. The first phase identified several practical barriers to combining multiple datasets, such as difficulty in reconciling data to projects and ensuring consistent entity identification, which was exacerbated by pseudonymisation. Several recommendations arise from this first phase including ensuring that all data is associated with projects and that processes should be transitioned to digital-first systems, not merely replicate their paper-based predecessors.

The second phase found that timesheet data could be processed and features engineered that were suitable for linear regression, AutoML and artificial neural network models. These investigations found that the modelling had some predictive power, but it is recommended that further investigations into richer datasets combined with alternative model types is required before the approach can be operationalised.

Elliott Hunston
May 2022