draft 1

Amy Cook

Tuesday, July 07, 2015

Originality Statement

'I hereby declare that this submission .... look up QUT'

Signed ................. Date....................

# Abstract

# Contents

# List of Figures

# List of Tables

# Glossary

B2B- Business to Business

# Acknowledgements

# Publications

# Introduction - 2 pages

* vague interesting few sentences - context

Predicting how long it will take a person to do something is a notoriously tricky task. People make mental calculations of these sort every day, guessing how long it will take to drive to meet someone on time, how many tasks you can check off in a day, or how long it will take to cook a dish. These mental calculations can be particularly inaccurate if the individual has never performed the task before. These errors are trivial in everyday activities, but they can have more meaning for businesses that sell their time as their chief product. For example, consulting businesses give expert advice to other professionals, in exchange for a fee for the amount of time the constultant spent on the problem. The delivery of 'expert advice' is generally in the form of a document (which could be anything from a report to hundreds of drawings) and can range from an hours advice to thousands of hours work on a substantial project. Before a project commences, the client and consultant must agree on a fee for the expert advice, or a fee structure. The nature of how a fee is structured can be as creative as the engaged parties wish, however a couple of simple examples are as follows: the client agrees to pay for the consultants time by the hour until the task is completed, or the consultant provides a fixed fee to complete the project in full, regardless of their hours spent. The nature of the fee structure brings an element of financial risk to one party which must be mitigated.

This thesis focuses on the risk a consulting company undertakes when setting a fixed fee for a project before their role has commenced. Clients commonly ask for competitive fixed quotes from several consultants before settling with one or negotiating further. The way a consulting manager decides on their fixed fee varies greatly from industry to industry and company to company. Typically, consulting manager has extensive experience in the type of project he is quoting and can use a combination of intuition, ratios of the entire project cost, and looking up a couple of past similar projects. Such a method often results in a wide range of profitabilities for projects depending on the ability of the manager to predict the amount of time the project will demand from the consulting team. This thesis examines a case study consulting company that has recorded internal historical project performance data since 2002. The study tests whether a statistical predictive model based on past project data can predict the profitability of a project based on characteristics available at the beginning of a project. This thesis asks further whether the prediction can impact a consulting manager's determination of a fixed fee.

# Chapter 1 Problem Description

In this chapter the research motivations are explained along with the goals of the project and the specific research hypothesis. Finally the contributions of this body of work are outlined along with the structure of the thesis.

## 1.1 Research Motivations - 4.5 pages

description of problem:

Project decision makers have struggled with forecasting project costs accurately for decades, across many industries. A study on large-scale infrastructure projects over the past seventy years revealed that cost forecasts consistently underestimated the cost of rail projects by 44.7%, the cost of bridge and tunnel projects by 33.8% and the cost of road projects by 20.4% (Flyvbjerg, 2006). A study across 1471 IT projects showed that 27% of projects ran over budget, and one in 6 of those projects were more than 200% over budget on average (Flyvbjerg and Budzier, 2011). These statistics cast a concerning glimpse at the financial story playing out behind the scenes of these projects which can cause job losses, ruin private businesses, and churn through government balancesheets. Complex long term projects are necessary to produce railways, houses, websites, and movies. This research focuses on the financial risk that is transferred to consulting companies who are enlisted to complete portions of these complex projects. The risk is transferred by requiring a fixed flat fee from the consultant, meaning they commit to complete their role for a nominated price. In a similar way to infrastructure and IT projects, consultants are susceptible to underestimating the amount of their time that will be spent in their role, and must wear the cost of extra employee hours spent. For privately owned businesses, mitigating this risk would increase the chances of the business surviving as well as growing and succeeding.

The ramifications of unprofitable projects in a private business are multifold. Employee morale deteoriates and employees may view themselves as incompetent if they are responsible for the project's delivery. Employees also experience stress attempting to complete the project within a disappearing budget and may produce lower quality work in exchange for speed. This is unfortunate if in reality, the project fee was under estimated. Unprofitable projects also limit a business' opportunity to invest money into marketing or staff training, which would improve a business' standing. The financial risk taken on fixed flat fee projects may also discourage business' from taking different risks with better potential pay-offs, such as expanding the business into a new area. Logically, predicting fixed fee project costs more accurately could positively affect a business' growth opportunities and staff morale.

Why its not easily solvable:

The challenge for consultants to reduce the risk of a project going over budget, in terms of hours spent, is not easily solved. Complex projects are always different from each other, and similar work may take significantly different amounts of time. For example, an ad filmed in a remote location for chocolate will cost significantly more than an ad filmed on an exisitng set, even though the product structure is almost identical. Therefore, a fixed product price, which you could apply to products such as appliances, is inappropriate for film projects. One option is charging for consultants' time and the hiring of materials by the hour; a zero risk fee structure. However in many industries this is not competitive enough. Other consultancies may be willing to offer a fixed flat fee for the same work, reducing financial risk for the client, and luring the client to them. Another fee strategy may be to track worker hours, and stop work once the consulting fee has been depleted. Then, negotiate further fees, or variations, with the client and persuade them to accept. This also satisfactorily reduces consultant financial risk, and in some cases, depending on the nature of the work or uncontrollable changes, this method can be effective. However, if the project has been following a predictable process but the hours are nevertheless inflating beyond expectation (a common situation), consultants may prefer to wear losses or endure very marginable profits to preserve the relationship with their client. Consultant's may believe it is in their best interest to maintain a reliable, trustworthy reputation in their industry over creating financial friction. Clearly, there is no straightforward solution to reducing the financial risk taken on by consultant's who offer fixed fees in complex projects. In some indusries, fixed flat fee structures are expected and business' must conform to survive.

current practise:

Limited research is available on the cost estimation techniques performed by industry for complex projects, however two surveys were performed in 2000: one on construction projects, the other on IT projects. The construction cost estimation study by Akintoye and Fitzgerald (2000) surveyed 84 UK construction contractors, ranging from small to medium to large, about their estimating practises. They found that the main method for cost estimation was breaking the project into detailed parts and adding up the cost of each item. The next two most popular methods were 'comparison with similar projects based on documented facts', and 'comparison with similar projects based on personal experience'. These can all be classified as experience based models (Akintoye and Fitzgerald 2000). A survey by Moores and Edwards in 1992 of 54 software developing companies found that detailed project planning tools were used by most companies as opposed to cost estimation tools, suggesting that projects were priced based on an analysis of a detailed breakdown of tasks within a project. In the same survey, 91% of software companies cited cost estimation as a problem. Although limited recent literature is available on industry cost estimation practises, historically it can be observed that for complex construction and software projects, cost predictions are chiefly made with a mixture of experience, project comparisons, and an analysis of the project details and tasks. It is also acknowledged that cost estimation is a problem and hypothesised that inexperience and lack of time are chief contributors (Akintoye and Fitzgerald 2000).

theory about problem with current practise:

Lovallo and Kahneman (2003) tackled the psychology behind the high failure rates of executives in predicting costs of projects such as manufacturing plant construction, mergers and acquisitions, large infrastructure and software development. Their theory stems from Kahneman's work on decision making that won him the nobel prize for economics in 2002. His research argues that humans' natural optimistic view of their own skills leads to consistent underestimation of the time and risks involved in a project. A manager optimistically sees challenges in a project as something that can be overcome by the team's high skill level, and downplay or ignore the risk of problems that are out of the team's control. It is for this reason that it does not matter if the project is broken down to the highest level of detail for cost prediction, all complex projects are at risk of encountering a multitude of problems that the manager could never foresee. Each problem has a low chance of occurring, but in combination the risk is much greater (Lovallo and Kahneman 2003). Lovallo and Kahneman (2003) call the practise of analysing a project based on project details and its unique complexities the 'inside view'. Research has shown that if people are forced to make predictions about their skills after being exposed to an 'outside view', their predictions are significantly more accurate. This outside view can be applied to complex projects, and involves ignoring the details of the current project and instead analysing the outcomes of several projects of a similar description. Mobilising the outside view in this way is called reference class forecasting (Lovallo and Kahneman 2003). One method Lovallo and Kahneman (2003) recommend is to obtain correlation statistics from past similar projects - the correlation between the forecast cost and the actual cost. The correlation for the current project can then be estimated via a statistical model which is used to adjust the forecasted cost made by detailed analysis (the inside view) (Flyvbjerg 2011). The traditional way to think about a project is to focus on the project details. Even though gathering data from similar projects could significantly improve a cost prediction, it is rarely thought of by managers (Flyvbjerg 2011).

The implementation of reference class forecasting began in project management for the first time in 2004, was endorsed by the American Planning Association in 2005, and is now used in some governments and private companies in Europe, South Africa, and Australia. An example of the type of output from a reference class statistical model is a plot showing the relationship between the acceptable chance of cost overrun and the required uplift to the original forecasted cost (Flyvbjerg 2011).



The visualisation presents a powerful communication tool for influencing decision makers and improving forecast cost accuracy, however access to credible data to a sufficient number of projects can be a challenge (Flyvbjerg 2011). Overall, the theory of optimism bias is a credible explanation for the problem of persistant underestimation of project costs. The accompanying solution of reference class forecasting presents a promising solution given enough accurate data.

previous research:

The theory of reference class forecasting was first clearly published in 2003, however the idea of predicting project costs from previous data has been the subject of research for many years before that and continues to this day. Although research results have often been promising, industry uptake of the idea is not yet established.

### 1.1.1 Overview of Past Reference Class Forecasting Methods and relevant Industry Uptake

In the following section, an overview of the historical attempts to integrate reference class forecasting will be presented followed by published theories for the lack of industry uptake.

***need to really do more research on this!!***

#### 1.1.1.1 Existing Models

A substantial amount of research on project cost prediction has been dedicated to the fields of software development and construction projects in particular. These analyses have generally used project information from 15 to 20 large projects. A single dataset may consist of projects from around the world, and each project a different client and project team. The mathematical methods used to predict project costs vary from statistical regressions to deep-learning neural networks (**???**)(Kim, An, and Kang 2004). Most studies report neural networks are more accurate than regression, and generally the studies report positive predictive powers of their developed models.

A McKinsey study outlined a case in the film industry where a motion picture company used reference class forecasting of movie project success (cases were weighted by similarity) to decide which movies should be heavily promoted. This model improved forecasts on financial return by 135% relative to single project analogies (**???**).

Statistical research projects which trial reference forecasting have shown promising results in IT, construction, and film projects. However, there is little evidence that the methods have successfully transitioned into industry.

#### 1.1.1.2 Industry Uptake of Existing Models and Current Practises

Surveys on current industry practise for cost estimation are scarce, however, a survey of the construction industry in the UK in 2000 found that a manager's experience and intuition is still the dominant method for construction cost calculation (Akintoye and Fitzgerald 2000). A study on the IT industry by M. Jorgensen et al. (2007) cited that the parametric models were not comprehensive enough to be used in industry and also found that expert judgment was not mobilised enough in the process of developing these tools.

Although reference class models have not gained traction in either the software or construction industry (where most of the research to date has focussed), it is worth outlining their current practises in cost estimation and the reasons behind these.

In the construction industry, infrastructure and building projects follow a traditional and well established process, where contractual norms have developed over many decades (**???**). The industry is so large, broad and well established that it has inertia in the way that projects are estimated, and contractually bound. The contractor (builder) first reviews the drawings and if the plans are not detailed enough, they offer a temporary budget price. Once they feel that the drawings are sufficiently detailed, they will offer a fixed price (**???**). However, as the job progresses, if any work is added that was not in the detailed drawings, they maintain the right to charge a variation (**???**). This method is clear and variations are easily justifiable to the client, however cost, labour and time estimates are still regularly under calculated, even from detailed drawings. And for large infrastructure projects, small percentage miscalculations can be exceedingly expensive.

Contractor's practise of pricing from detailed drawings and charging for variations does not translate to other consultants in the construction industry, such as engineers and architects. Consultants must devlop the detailed plans for the contractors, and must therefore offer fixed price contracts based on drawings with little detail. It is much harder for a consultant to charge a variation for a portion of their time spent on something that was not included in the original simple plans. This is despite the real and common instance of budget blow outs due to overspent time (**???**). One method to provide opportunity for a variation is stating a minimum percentage fee of the final building cost, which can overrule the original fixed price contract. If a consultant has spent time beyond their fixed price budget,this may account for some or in the best case, all, of their overspent time.

Construction projects generally are completed more on time and on budget when compared to IT projects as the greatest cost is the project is building materials (**???**). Unlike time, building materials can be accurately calculated from detailed design drawings. Again, this omits data from consultants who are part of the building industry but charge based on their time only.

In contrast to construction, IT projects are a relatively new practise where alternative contractural arangements have been trialled. Through experimentation with contracts, the Agile method was developed. In summary, the method treats both cost and time as fixed quantities for a project, and if any change or additional work can only be accommodated if another, less important requirement is excluded (**???**). Furthermore, the project outcomes are continually adjusted ad revised based on frequent progress review points that assesses unexpected problems. This agile attitude has become ingrained in many IT practises and this flexible upfront arrangement with clients often forges better long term relationships [Cockburn2001].

It is not completely clear why industry uptake of cost estimation tools using the outside view has been weak, and it may be the case that in modern industries such as in the IT industry, problems associated with cost estimation can be innovatively designed out of the contract. Possible reasons that parametric reference class forecasting models developed by researchers have not been incorporated into industry practise include:

* lack of appropriate data
  + if a project is unique, such as a national infrastructure project, it many not feel appropriate to gather sufficient data points from other countries and time periods.
* lack of understanding from industry decision makers - the prediction process was not explained well in marketing of the tools
* models were not developed in collaboration with industry
* does not feel intuitive price from an 'outside view' of a project which disregards small project details
* insufficient time allocated to pricing that makes change burdensome

In fact, a study by Moores and Edwards (1992) concluded that a lack of framework to support the outside view model use and a failure in marketing were the two most common explanations for lack of uptake of cost estimating tools which use statistical techniques in predicting project costs. This is unfortunate as almost all research indicates an outside view analysis of a job substantially improves cost estimation accuracy for fixed price projects.

### 1.1.2 Case Study

This research will explore the potential for reference class forecasting to improve cost estimation in the context of an engineering consulting company in the construction industry. In this field, labour is the chief cost and the long traditional history of this industry mandates fixed price projects as the norm. In recent years, the size of losses from unprofitable projects in the case study could equal up to 25-30% of the profits from profitable projects. This demonstrates substantial room for improvement and the possiblity of multitude positive flow on effects.

Cost estimation is currently performed by first carefully reviewing preliminary drawings for a project where time costs will be calculated using a manager's personal experience with similar projects. This will often be cross-checked with a value based off a percentage of the estimated final cost of construction (generally the client budget). No formal mathematical method of comparison to similar projects is performed. This is the case for a number of possible reasons:

* lack of time to interrogate the CRM
* clunky data availability offered by the CRM
* lack of awareness of projects performed by other managers that may have been similar.

Reference class forecasting in this study differs from previous research, in that statistical analysis will be performed on internal past company project data that has been progressively recorded in the company's Customer Relationship Management (CRM) software which was purchased. This presents an opportunity to evaluate reference class forecasting in a unique situation, where the 'reference' projects are internal and over two thousand cases are available.

The CRM data currently stands as an untapped source of information within the case study organisation. It stores a rich variety of data including:

* employee timesheet hours with dates
* other project costs (taxis, printing)
* client information/characteristics
* client identification code
* invoiced amounts for each project and dates
* employee costs
* employee charge out rates
* project description

A statistical model predicting a measure of project financial success will be developed using the available project variables and trialling numerous machine learning algorithms and statistical methods. The company uses well established CRM software which readily provides simple output statistics such as project time-cost vs invoiced amount summaries. However, the analytical capabilities were limited to:

* simple scatter plots and bar charts of the raw data
* summaries such as overall hours spent vs invoiced amount for each project

The company has a wealth of data but limited means to extract insight. This study will exploit the CRM data by performing more sophisticated statistical analysis with the intention of building a predictive model for cost estimation. The benefits of this study in comparison to previous cases are that thousands of past data points are available, when previous studies used on average 15-20 cases and at most 170. This improves the potential for accurate predictions. If successful, the model also has higher potential for managerial uptake as the algorithm will be directly built and trained on company data. Managers can interrogate 'similar' data points and relate to the actual projects or speak to a colleague who was involved. This has the potential to give the 'outside view' stronger influence on the final decision.

### 1.1.3 Case Study Limitations

The case study provides a framework to test the value of a consulting company's CRM data in assisting project cost estimates. However, it is important to be aware of limitations when reviewing outcomes and framing the problem. The most obvious limitation is that we are disection a single company in a specific industry, which limits the ability to state conclusions could apply to a broad range of cases. This can be overcome with future studies.

The data set from this case study does not include information that would improve project cost predictions. For example, managers commonly price jobs based on the budget for the entire project, of which the consultant is a minor part. The manager might typically moderate his fee by say 1% of the entire project cost. The value of the entire project has typically not been recorded although there is a space in the project software form for this value. It would be even more beneficial for managers to record the expected total project cost at the beginning of a project versus the conclusion of a completed construction project. Another piece of data that was not recorded was external factors influencing the fixed price quote including competition from similar firms that the client played against the case study company.

In addition to this data that was not typically entered by workers, the dataset also excludes detailed information about the projects which are crucial to the final pricing. This includes details such as preliminary drawings describing the project (square meterage, number of storeys), and project summaries provided by the client to enable a quote. This kind of detailed information will never be recorded in a CRM and in many ways contradicts the purpose of the 'outside view' which intends to present similar projects without being over-influenced by the finer details of a project. Given the lack of detailed information, it is not reasonable for an algorithm to predict the specifc fee of a project. It is much more effective to attempt predict the profitability of a project (ie return per dollar) based on the historical data. This profitability prediction relates to the profitability of the project if the fee was generated using the established method the company used in its historical cases.

Finally, CRM data is entered by employees manually as time passes, jobs are invoiced, and projects are finally completed. Because the data is input by humans, it is susceptible to input error. Many of errors were detectable during data cleaning, however it is possible that undetectable human errors exist in the data.

Although this case study presents an opportunity to test the predictive power held within the organisations' existing CRM data, limitations exist which must be understood and built into the framework for predictions. These include missing data that could have been entered for each project, but historically has not been required, missing data that is very detailed and will never be available, and the potential for a small degree of human error to be embedded within the data.

### 1.1.4 Research Motivation Summary

Fixed price cost estimation for complex projects is a difficult task and when performed incorrectly, can have devastating affects on businesses and governments. Many theories have delved into human nature and our tendency to optimistically assess our capabilities with respect to a given task. In this case, the optimism often results in negative business consequences affecting employee morale and the capacity of a company to flourish and grow. Over the past few decades, significant research has been dedicated to creating predictive models that present the 'outside view' of a project, by statistically comparing a new project to numerous similar projects and their characteristics. It has been shown that these models improve the cost estimation, however industry uptake has not been successful. It has proven to be a challenging task to persuade managers to use this so called reference class models in practise and remove themselves from additions based on intricate details of the completed project. This may be due to a number of reasons including lack of understanding, time, and lack of collaboration with industry in development of these models. This research examines the potential for a company's internally generated CRM data to be used in creating a reference class forecasting model. This has the potential to influence the managers' decisions in the company more however its limitations must be understood in order to create a good model.

## 2.0 Thesis aims

For the reasons explained above, the aim of this research is as follows:

**General Aim** The aim of this project is to use statistical techniques to model the profitability of projects for consulting businesses using their internal CRM data. Research will focus on a case study Engineering consulting company that offers their expert advice (in the currency of time) to business clients. The project outcomes are intended to assist the business in predicting project costs before project engagement and other business analyses such as client analysis. Several statistical and machine learning techniques will be tested, compared and refined.

**Hypothesis 1** A statistical or machine learning model based on historical project data can predict the profitability of a new project.

**Hypothesis 2** The predictive model built from Hypothesis 1 can be proven to have a positive impact on the bottom line of the case study business.

## 1.2 Thesis contributions - Gaps in literature

* a gap in the study of predicting internal performance in businesses with business clients. The literature has assessed performance of movies, customer analytics but rarely employee performance or business clients. This is particularly relevant to consulting companies that tackle complicated discrete projects with set timelines. Statistical and machine learning models to predict the correct fee for projects according to project type, client, and other characteristics could reduce risk for a range of consulting businesses in setting initial fees.
* A gap in machine learning predictive models which cater for variables with large amounts of missing data. the nature of consulting jobs is quite complex and historically, not all information about a project is possible to obtain. Variables with missing data are alternatively treated as optional bonus variables in a multi-step model as opposed to missing information which must be imputed or deletion of the entire case that contains missing information.
* the application of clustering methods to business to business decisions.

## 1.3 Thesis structure

lit review Background info present the case study Present the models you've developed summary of findings, implications for use within industry, limitations

# Chapter 2 Literature Review - 25 pages for pHd

This chapter provides an overview of the literature available about current business to business (B2B) cost estimation methods, use cases of CRM data, as well as machine learning algorithms and statistical business prediction models. The most prominent statistical and machine-learning models developed for business applications shall be reviewed along with models applied to the most closely related business cases to this project. Gaps in past research will also be highlighted.

## 2.1 Project Cost Estimation Methods

This project will tackle the problem of estimating fixed fees for complex consulting projects. It is therefore imperitive to understand what research has been performed to date and how this project contributes to the body of work. The task of estimating fixed fees can be likened to the terms 'cost estimation' or 'effort estimation' in the literature. The bulk of research to date has been performed with project data in either the construction industry or software development. There is also a body of research on predicting returns on movies in the film industry. The research into cost estimation in each of these three industries is reviewed below, detailing different methods and outcomes. Finally a gap in the literature will be highlighted which this research aims to fill.

### 2.1.1 Cost Estimation in the Construction Industry

Research into cost estimation in the construction industry chiefly focuses on predicting the final building cost of construction. Various studies have surveyed popular methods for cost estimation while others have investigated new methods such as building information modelling (BIM) analyisis, case based reasoning, and predictive algorithms. The cost estimation methods of detailed analysis (inside view), case based reasoning, and predictive algorithms shall be reviewed.

**Detailed Analysis**

Detailed analysis refers to the process of an engineer or builder in the construction team carefully reviewing construction drawings and sum the cost of materials, labour, machinery hire, overheads, profit etc. (Akintoye and Fitzgerald 2000). Akintoye and Fitzgerald (2000)'s study based in the UK found that this is the most popular method for pricing amongst contractors. The second most popular method was to perform an additional manual comparison with similar projects and make judgments based on personal experience. Note that this does not involve a mathematical model.

(**???**) discusses how much detailed cost estimates for cosntruction projects can vary from estimator (engineer) to estimator, which contributes to the lack of accuracy in predciting final cost of a project. He argues that an engineer's expertise is not documented or measured in any way and therefore their expertise and therefore estimate is prone to subjectivity. (**???**) theorise that final cost can be subject to so many different unpredictable parameters, such as weather, unexpected ground conditions, duration or sub-contractor issues. It is therefore almost impossible to achieve an accurate cost prediction manually (**???**).

Groundbreaking methods into calculating construction costs by (**???**) mobilise the data from building information modelling, although this method is still technically a detailed analysis of the costs. The idea is to quantify the cost of building a structure directly from a three dimensional model created by the design consultants such as engineers and architects. (**???**)'s first trial was to automate the cost calculation of structure composed of reinforced concrete only. His chief development was to program his algorithm to intelligently establish construction techniques for each element as this is not provided by architectural or engineering models. From the construction technique and material, a cost can be determined for each element. This was successful and Ma aim's to further his work so that more construction information can be intelligently obtained from BIM models. A system like Ma's would definitely accelerate the cost estimation process and reduce human error, however it is still an 'inside view' and could be prone to innaccurate estimates in a similar way to a detailed analysis by a person. The detailed view does not 'step back' and take into account set backs in construction that may have affected similar projects in the past. There is potential for this to be built into the same system but this is still in early stages of development.

Survey information on the latest techniques for cost estimation is limited, however what is available indicates that cost estimation through detailed analysis (inside view) remains the most prevalent method. This is despite construction having a long history of projects that have run over time and over budget using the same cost estimation technique (**???**). There is even current research on how to automate this way of cost estimation using BIM. The method of detailed sums shall be compared to the results of studies using parametric algorithms and case based reasoning ('outside view') in the following sections.

**Statistical and Machine Learning Models**

As the problem of cost estimation in construction projects is well recognised, dozens of studies have built prediction models of these projects using machine learning and statistical models. The aim is to test whether a mathematical comparison of many similar projects, using meaningful descriptive variables, can predict the cost of construction. This section will first review the previous studies and most accurate algorithms and finally summarise industry uptake of the methods.

(**???**) completed a study on cost estimation techniques and research from 2004 to 2014. It was found that artificial neural networks and support vector machines were the most common machine learning techniques used. He highlighted that these two machine learning techniques deal well with uncertainty, however lack technical justification to the decision maker (known as a black box). Neural networks are also time consuming to train, and must be re-trained and re-tested with each additional piece of data (Kim, An, and Kang 2004). Nevertheless, neural networks and SVM's recieved a significant amount of attention in the 1990's for their ability to accurately predict construction costs with limited detailed information (**???**)(Kim, An, and Kang 2004).

(**???**) pioneered the application of boosted trees to cost estimation in construction projects. This is surprising given the rapid uptake of boosted trees over the past decade. In the study, data from 234 school building construction projects in Korea was used. Boosted trees were compared to neural networks, which are one of the most popular algorithms to apply to the problem. Boosted trees were found to predict costs more slightly accurately than neural networks, but not significantly (p value < 0.5). Even though the two algorithms performed similarly, the output from boosted tree models provide insight into the structure of the model such as variable importance and variable partial dependence plots. This additional insight may mean boosted trees are preferable over neural networks because of their potential to engage decision makers.

A notable amount of literature has studied the predictive power of multiple linear regression in the construction cost estimation problem. Often, regression is the only model assessed, without comparison to other methods such as neural networks, which first started appearing in literature in the 1990's (Kim, An, and Kang 2004). However, even post-1990, many studies focused on regression only. This may be because the technique is straightforward, easy to use and widely available in statistical packages (**???**). Out of those that have compared regression to neural networks, the studies have shown that neural networks outperform regression, however other studies show they are approximately equal (Kim, An, and Kang 2004)(**???**). (**???**) found that neural networks outperformed multiple linear regression models, however regression models could be performed first to determine variable importance and condense the number of variables. It can be argued that boosted trees could perform as well as neural networks while providing additional insights (even more so than regression models) such as variable importance and partial dependence of variables against cost.

Despite neural networks typically outperforming regression models, it is intersting to note some of the outcomes of regression studies in cost estimation. Different studies tend to produce a wide range of variables that contribute to different degrees in the models. This is likey due to the nature of construction projects which are multidisciplinary and involve a range of different parties such as the clients, consultants, contractors and suppliers, each of which can have a range of characteristics and degrees of input (**???**). (**???**)'s study on 87 building projects in Singapore included special, complex projects. Therefore, variables such as contractor's specialised skills, who the client was (public vs private), the client level of experience, and the contractor's financial management highly influenced final cost of the project. Other studies found variables such as project complexity, duration, team experience, information availablility, site requirements, and labour climate to contribute most (Akintoye and Fitzgerald 2000)(**???**)(**???**). As these studies obtained information from questionaires, the available variables also largely depended on the questions in the questionaires, and therefore each study produces a range of different important variables. Despite the differences in qualitative outcomes, the studies all generally report their mean absolute error or a similar metric from the tested models. This is encouraging, however there is a gap in the literature in how to translate the improved errors from the model into an argument for industry to uptake the model as an improvement on their bottom line.

Surveys into construction industry uptake of these algorithmic models are sparse, however a survey by Akintoye and Fitzgerald (2000) showed that parametric estimating methods have not been adopted by contractors, despite the research into this area over 20 years prior. The majority of contractors use experience based methods. Akintoye and Fitzgerald (2000) hypothesises that this could be due to a lack of familiarity and knowledge about the techniques, doubts whether these techniques are applicable to the construction industry, and the availability of sound data to ensure confidence. This presents a clear gap in the literature on how to analyse the models in terms of the goals and limitations that a contractor may experience.

The literature on cost estimation for building projects do not break down the cost estimation task into smaller components, even though a construction project is completed by dozens of consultants and sub-contractors. Many of the consultants, such as Engineers and Architects, produce their own cost estimates for their effort that contributes to the project. Their contracts are often fixed price, with minimal room for movement, which means they wear the risk of over-spending to meet fixed cost contracts (**???**). Furthermore, the existing studies collect data from a wide range of projects that have certain characteristics in common, such as country, or building type and information is obtained through surveys. This leaves a gap in research exploring a single contractor's history of projects, which would be available in their company database.

**case based reasoning**

Case based reasoning (CBR) is a method of estimating by using the results of similar cases. The idea is that in the case of a new project, similar projects are chosen, before generalisations about the data are made. This contrasts to algorithmic models that are built from the entire set of data (**???**). Research into the application of case based reasoning to construction cost estimation began in the 1980's (Kim, An, and Kang 2004). The general procedure involves:

1. Storing a collection of projects with key variable values
2. Once a new project or case arrives, similar cases are retrieved. This can be achieved either algorithmically with a distance function applied to the variables, or manually by users reviewing past cases. The nearest neighbour algorithm has been used to find the distance between cases in previous studies. It is programmed to calculate the euclidean distance in n-dimensional space between cases, where each variable is a dimension (**???**).
3. The cost of the new project is estimated by extrapolating characteristics of the similar cases to the new case. This can also be done either algorithmically or manually by an experienced decision maker.

A study by Kim, An, and Kang (2004) compared CBR to neural networks and regression in the construction cost estimation problem. For CBR, the ESTEEM software package was used which algorithmically calculated the similarity of variables, weighted the variables using gradient descent, and deduced the cost using the most similar cases (the specific method for this last step was not explained). The discussion compared mean absolute error rates of the three methods and showed that CBR was more accurate than regression, but less accurate than neural networks in cost estimation. Despite neural networks outperforming CBR, CBR maintained significant advantages over neural networks which have been referenced in other studies as well.

Advantages of CBR include algorithmic efficiency as well as user engagement. In Kim, An, and Kang (2004)'s study, it was noted that CBR models were simple to update with new data in comparison to neural networks, which must be slowly re-trained and re-tuned. New cases also did not need to have every variable complete which can occur often in real, messy, data. Neural networks on the other hand require complete data. Another significant advantage of CBR is the ability for users to review chosen similar cases and make sense of the predicted cost, as opposed to neural networks which are a black box (Kim, An, and Kang 2004)(**???**)(**???**). A disadvantage of CBR is the accuracy can be highly dependent on the number of selected cases (**???**). Overall, CBR may provide prediction accuracies between regression and neural networks but its results can be justified by the user.

It can be argued that CBR is a systematic method of expert judgment, where the decision maker manually compares similar projects from his or her experience (**???**). CBR that is performed manually through personal experience, without the assistance of algorithms, was the second most popular method for construction estimation in Akintoye and Fitzgerald (2000)'s survey in the UK. This indicates it is quite an intuitive method for decision maker's to cross check their detailed cost estimate, although an algorithmic version of CBR has not been taken up by industry.

**summary**

In summary, research into cost estimation in the construction industry has been undertaken using a variety of methods ranging from detailed manual analysis, algorithmic models, and case based reasoning. Neural networks generally outperformed all other systematic methods, such as regression and case based reasoning, but had a significant disadvantage of being a black box. Alternative methods such as boosted trees have the potential to perform as well as neural networks and provide insight into the structure of the algorithm. This is true for CBR as well, although it has not been shown to predict as accurately as neural networks. There exists a gap in the research in ways to intelligently combine CBR with machine learning methods that could predict well, provide insight to the structure, and engage the user in reviewing similar projects. Also, as stated previously, there exists a gap in the application of cost estimating in the construction industry to smaller components of the project, such as consulting and contracting companies who each face a similar problem in estimating their fixed fee. These companies could mobilise information in their internal project databases, as opposed to available literature which surveys dozens of companies involved in selected construction projects.

### 2.1.1 Effort Estimation in the IT Industry

**you are here**

## 2.2 Use cases of CRM data

* no idea
* gap - my project obviously

## 2.3 Statistical and machine learning methods applied to business problems

### 2.3.1 Most simple and frequent models

* anova
* linear regression
  + pros
  + cons
    - do these acknowledge their data is not normally distributed?

### 2.3.2 Complex models

* SVM, Neural networks, random forest
* data must not be normal
* most successful?

### 2.3.3 ensemble methods

* papers
* netflix?
* multi step to deal with missing data

### 2.3.4 similar applications

* Employee churn - anything similar?
* stepped method business application?
* Gap

\*\*

## 2.4 Conclusion

summary of \* how are B2B decisions currently made \* what are the most popular/effective methods \* which ensemble methods may suit our missing data situation \* Gap in literature/contribution

# Chapter 3 Introduction of Case study

## 3.1 Obtaining dat

### 3.1.1 How data was recorded

* years of employees entering hours each day
* project managers or admin entering client details

### 3.1.2 How data was extracted - CRM

* queried direct from CRM
* data de-identification

## 3.2 Summary of variables

* categorical
* numerical
* significant cleaning
  + cost sometimes entered in as hours. ie $24 entered as 24 hours
* engineered variables
  + grep through project names to create categories - categories developed with case study manager
  + return per dollar = (invoiced - cost)/cost
  + hours analysed to produce: employee position that performed majority of the hours
    - % of hours completed by majority position
  + timespan based on timesheet entries as well as number of days with hours recorded
  + number of workers on project based on timesheet entries
  + client characteristics such as mean total amount invoiced for a repeat client
* amount of missingness

## 3.3 Variable selection

* combination of cforest, anova, random forest

# Chapter ? Method

## Regression

\* tried predicting return per dollar as a numeric outcome.  
 \* proved very difficult. Find error rate compared to just guessing the mean. Not far off.

## Bayesian Network

* could not handle *any* numeric variables, had to discretise all numeric vars
  + used normalised numeric variables
  + I did this by making a hierarchical dendrogram to visualise the clusters
  + then I chose the number of clusters, drew the red rectangle
  + summarised the clusters to find max and min values for the variable within each cluster. Rounded this value up
  + manually discretised variable using rounded values
* first trialled complete cases of number of users, Discipline, percent pro, Business, client code, job detail, majority position, timespan, invoiced amount, percent hours by majority position, and return per dollar. This gave 907 complete cases to work with.

*using Genie Smile.*  turned return per dollar into profit/loss. This gave 182 loss cases and 725 profit cases. A good distribution for working with bayesian networks. Ie not too biased \* a good, simple problem to start with

* establish a baseline. Using ROC, area under the curve method
  + to test our classifiers we need to beat random chance. Ie AUC = 0.5
  + will also establish the AUC for using one variable only! See table below: \*Note all using 10 fold cross validation

|  |  |  |
| --- | --- | --- |
| Single Variable | AUC from ROC chart | Num losses correct |
| Random chance | 0.5 | ? |
| percent pro | 0.618 | 0 |
| percent maj pos | 0.616 | 0 |
| number of users | 0.661 | 0 |
| timespan | 0.600 | 0 |
| amount invoiced | 0.535 | 0 |
| Discipline | 0.514 | 0 |
| Majority pos | 0.605 | 3 |
| Job detail | 0.605 | 1 |
| client code | 0.567 | 25 |
| client Business | 0.524 | 5 |

* try using the following variables:
* number of users
* Discipline
* percent hours by a professional
* position that completed the majority of hours
* timespan
* amount invoiced
* percent of hours completed by majority position

The results are as follows:

AUC = 0.705 Number of losses predicted correctly: 124

This is our 'base' model. Now try adding one complex variable at a time:

|  |  |  |  |
| --- | --- | --- | --- |
| Added Variable | AUC from ROC chart | Num losses correct (tp) | Num losses wrong (fp) |
| none/base | 0.710 | 122 | 272 |
| client code | 0.624 | 171 | 528 |
| Busines | 0.667 | 144 | 381 |
| job detail | 0.653 | 150 | 438 |

## Boosted Trees

### Advantages

\* Wonderful in that NA's don't interfere with prediciton process  
 \* explain..  
 \* can comfortably include the entire dataset without having to worry about imputing

tuned which variables to include: \* exclude JD.Second, Business, code.client, code.contact \* tuned parameters: \* shrinkage = 0.001 \* n.trees = 4000 \* interaction.depth = 3 \* min.nobs = 10 \* manage to get AUC up to 0.785!!

can we improve this by bringing huge variables back in a smarter way?

Akintoye, Akintola, and Eamon Fitzgerald. 2000. “A Survey of Current Cost Estimating Practices in the UK.” Journal Article. *Construction Management and Economics* 18 (2): 161–72. doi:[10.1080/014461900370799](http://dx.doi.org/10.1080/014461900370799).

Flyvbjerg, Bent. 2011. “Over Budget, over Time, over and over Again: Managing Major Projects.” Journal Article.

Jorgensen, M., Magne Jorgensen, M. Shepperd, and Martin Shepperd. 2007. “A Systematic Review of Software Development Cost Estimation Studies.” Journal Article. *IEEE Transactions on Software Engineering* 32; 33 (1): 33–53. doi:[10.1109/TSE.2007.256943](http://dx.doi.org/10.1109/TSE.2007.256943).

Kim, Gwang-Hee, Sung-Hoon An, and Kyung-In Kang. 2004. “Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning.” Journal Article. *Building and Environment* 39 (10): 1235–42. doi:[10.1016/j.buildenv.2004.02.013](http://dx.doi.org/10.1016/j.buildenv.2004.02.013).

Lovallo, Dan, and Daniel Kahneman. 2003. “Delusions of Success: How Optimism Undermines Executives’ Decisions.” Generic. HARVARD BUSINESS SCHOOL PUBLISHING CORPORATION. <http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMw3V3Pb9MwFLYQQxMSEgxYKAPJBwQHlCqO61-TdoC2aNImcdgmcYscxwGktoMlQfvzeXYSJy1M4swxjZuq_uz3vs95PxCi6TSJd2yCAU-T53pmjbIgoblMCqGYyEtpEpoWs-3zt9Bqb_jsfwB-YVdN1Ye3XTS-IaLT_a553GcwEGvXFsO3O1q7kPf3y1trGmf0fDzFouu5U41Za99AKETJb79ROL_-pbtXOIthsZ3pbxvbna-2iexbJww0RKMGq9l5srHVlGS0OsTIBDI-cqakTczcrnO9439CVCCIZQ6ml751Vc_XxXdTn9hNfHUBDhZ4q1cFH5d3utk_feqIVDxwOT5N9Vd24ZnE5RP0uC_pjT-02B2ge3bzFO33k_sMzQOE-LrEHYTHGADEPYB4ABAPAL7DAb7n6OrT8nJ-GnfNLuKvBChkzFTJcgnkVadGg1XkVllKiXYEDHYLYxKUK3imPCfGklIwA8KxYJQpSsAaSHqIHmmXFLGpffJkEaG9EhayjRyriOBfRmj_izpfyNOzeXt50F9OK5_hN_1ZRzC7fh_EfCpeICxEIg2fcebqtoHtzlOic26UYobKWZFMUOQmPXObor7RJiMJT0BSJGqCDnscsmK1yqgrlQfklMBXWlSyH209lQwoEgPdDnfetDCFO2lWpRmoVBDJqRBOq9a3NTxgZ1i_cOABY3jDfSdkYIhgzEueCSL_MmzeFcl3xSHql3f-6BF6OGyaV-h-fdPY1z5t-DdLboxJ>.

Moores, T. T., and J. S. Edwards. 1992. “Could Large UK Corporations and Computing Companies Use Software Cost Estimating Tools? - a Survey.” Journal Article. *European Journal of Information Systems* 1 (5): 311–20. doi:[10.1057/ejis.1992.3](http://dx.doi.org/10.1057/ejis.1992.3).