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 burn through government balancesheets. Complex long term projects are necessary to produce railways, houses, websites, and movies. Depending on the industry, the the financial risk of completing a project on budget is transferred to consulting companies who are enlisted to complete portions of a project. 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 perform 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. Clearly, 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. The reason is that 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 away. 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 low 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. The 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 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. This 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 (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, 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).

This method was implemented by the British government for a rail project which is still under completion

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.

* research: dozens of papers using statistical and machine learning to predict project costs - construc - IT - film costs?
* - reference class forecasting - pioneered in british government large gov infrastructure

Previous research has been done on the prediction of final costs of a complex project with varying success. In the prediction of IT software development, formal estimation methods have existed for many years, however expert judgment is still the dominant method and available evidence does not suggest the models improve on expert judgment (Jorgensen and Shepperd, 2007). Construction industry research has used models ranging from neural networks to case based reasoning to regression to predict final building costs. However, often expert judgment was not mobilised, and the models were not comprehensive enough to be used in industry. Manager's experience and intuition is still the dominant method for construction cost calculation.

Potential solution: \* stat analysis like reference class forecasting and building construction except analysis performed internally using the company's historical data on past projects.

case study:

This project undertakes a case study on a consulting company in the construction industry. Specifically, the

* in the case study (and is typical of smaller organisations??), the CRM is an untapped source of information storing
  + employee hours
  + client information/characteristics
  + invoicing
  + employee costs
  + employee charge out rates
  + project information
* currently the case study company uses a widely popular CRM, where they can output information but cannot perform modelling or analysis:
  + look up one job at a time
  + basic summaries of overall hours spent for the month

### 1.1.1 Limitations

* looking at a single case study company in a specific industry
* Managers commonly price jobs based on the budget for the entire project, of which the consultant is a minor part. An example of this is a consultant who typically charges 1% of the total project cost
  + this data has not been historically recorded. Can start being recorded and would be extremely useful
  + even better to have expected cost of entire project at start of project, as well as actual final cost of project.
* A fee is based off detailed information that is unique to each project, for example the size of a building in terms of storeys and square meterage. This detailed information is not available in the data, and is too complex and varied for an algorithm to digest. Therefore it is not reasonable for an algorithm to predict the fee of a project at this stage, but much more achievable to predict the profitability of a project (ie return per dollar) based on historical data. This profitability prediction relates to the profitability of the project if the fee is generated using the same method the company has used in its historical data.
* Although the aim is to predict the profitability of a project, the model aims to change managers' fee setting decisions if the algorithm predicts low profit or a loss based on historical data.
* external factors forcing price down in some cases - competiton. clients

## 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 decisions for consulting businesses. Research will focus on a case study Engineering consulting company that sells their expert advice (in the currency of time) to business clients. The business decisions include predicting project fees before project engagement and client analysis using existing invoicing and hourly employee-cost data. Several statistical and machine learning techniques will be tested, compared and refined.

**Hypothesis 1** k-means clustering methods can identify groups of similar projects with significant differences in profitability.

Unsupervised learning to identify clusters is a successful marketing technique used by retail companies to target groups of similar customers with similar habits and preferences. In this case of consulting companies, clusters may have multiple uses. New projects may have characteristics that slot into an 'unprofitable' cluster and the decision is clear to not engage in the project. Alternatively, if past projects and accompanying clients are found to fall into a very profitable cluster, consulting managers may focus services on these clients and reach out to these clients in particular for new projects.

**Hypothesis 2** A statistical or machine learning model based on historical project data can predict the profitability of a new project with sufficient confidence to influence a project manager's fee proposal.

## 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) decision strategy as well as methods in machine learning and statistical business prediction models. In particular, the literature on B2B decision strategy is centred around setting fixed consulting fees and methods of targeting new clients or projects. 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.

## 2.1 Business to Business Decision Strategy

### 2.1.1 Client and project analysis

* clustering
* how do businesses currently review their projects or clients

### 2.1.2 Fixed fees

* fixed fee problem in the IT consulting world - developed new techniques?
* managers currently go by gut feeling or look up a couple of past examples
* difficult to combine insights from similar jobs/clients

### 2.1.3 Gap

* no mobilisation of CRM data into sophisticated models in B2B world

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

### 2.2.1 Most simple and frequent models

* K means
* anova
* in construction journals, commonly applied linear regression
* do these acknowledge their data is not normally distributed?

### 2.2.2 Complex models

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

### 2.2.3 ensemble methods

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

### 2.2.4 similar applications

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

## 2.3 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

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.

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>.