Discover Insights from Data Analysis

Andrew Ho

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-118
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-118.html

May 15, 2015

CAPSTONE PROJECT

Discovering Insights from Data Analysis

Capstone Problem Statement, Capstone Project Strategy, Competitive Strategy, & Intellectual Property Protection by: Andrew Ho, Franklin Ma, Ives Huang, Sarmad Khalaf, & Sherry Chan

Technical Contribution: by Andrew Ho

May 4, 2015

Table of Contents

1. Capstone Project Problem Statement

What keeps you up at night? This is always the first question we ask small restaurant owners while interviewing them. No matter how many restaurant owners we interviewed, the answer was unchanged. Restaurant owners strive to save as much cost as possible to monetize their business. They emphasize labor cost as the major component in their overall cost structure.

However, after our observations for a couple of weeks, we discovered a surprising fact: In many restaurants, there is always a constant number of staff throughout the day. Yet the number of customers coming in fluctuate greatly with time. For example, there are on average two transactions made per minute at La Vals, the pizza place, around 12pm, while there is only one transaction made in three minutes at 3pm. Number of staff, however, remains eight at any time in the day. In other words, restaurant owners are either over or under staffed most of the time. While labor cost is a burden for all restaurants but surprisingly, nobody is solving this problem. That's why our capstone project is important. We are here to tackle staffing problem for restaurant so they would be able to focus wholeheartedly on making fantastic food. Our projects, if implemented successfully, would revolutionize how restaurant owners make staffing decisions.

Our capstone team helps restaurants reduce labor cost by generating a dynamic staffing schedule for them. We do that by training our software to predict how many people will visit at a specific time during the week, and match that to how many staff the restaurant should hire at that

shift. By leveraging our data driven service, staffing will become more efficient. Furthermore, we added behavior-modeling analysis, which would potentially generate better match between restaurants and staff. Our vision is to empower small business with technology so that they make the best decision for themselves. We believe everyone deserves to benefit from data to optimize their business with ease.

2. Capstone Project Strategy

2.1 The Vision

Intuit belongs to the financial software business, and they build the operating system for small businesses. Major players in this industry consist of Intuit, Square, Xero, Automatic Data Processing Inc., which all aim to make finance and payment easy for small business (Edwards, 2014). Our capstone team partners with Intuit to help small restaurants save labor cost. With our technology, we develop a staffing recommendation system through machine learning prediction and data optimization. We imagine our final deliverable as a value-added plug-in service for QuickBooks. Throughout our capstone project exploration, we aim to fuse our data creativity to not only expand our horizon but also help restaurants make better staffing decisions.

2.2 The Industry

There are several distinctive features in the financial software industry. First, data security is critical, and will become increasingly important as they collect more data. Therefore, public companies are spending more capital on protecting their data, and the trend is drawing

more attention from small businesses too (Barrett, 2014). In the near future, people won't use your product unless their data is extremely secured. In addition, clients need a high level of trust to perform financial transaction on the platform that makes network security and fraud detection of extreme importance. This creates a high entry barrier for new entrants to prove themselves as a reputable brand. Secondly, once customers decide to deploy a company's product, it's very unlikely they will switch to another venders' product due to high transactional cost and slow learning curve for small business owners (Higginbotham, 2013). In addition, the original vendor has all the users' data in hand and has a bigger advantage of providing valuable customized service leveraging data. Thirdly, clients in our industry usually lack formal business training. As a result, financial software needs to be user-friendly, simply intuitive, and with great visualization. Companies in this industry tend to offer free trial for customer for a short period and generate profit via subscription fee later. They work very closely with the users to understand their needs and to ensure their products create value for small business. To sum up, this is a unique industry with high entry barrier, but we can retain customers once we got them.

2.3 Competitive Landscape

Compared to other players in the industry providing software for transaction, accounting operations, and financial statement reviews (Edwards, 2015), our product is more customer centric, data driven, and analytic. Viewing ourselves as a QuickBooks value-added service, we have a huge competitive advantage under Intuit's brand. With competitive technology and a deep domain knowledge of customer, we will distinguish ourselves easily in the market by our analytic technology. Our technology will integrate well with QuickBooks and create a higher

entrant barrier. The following paragraphs will elaborate on technologies in this industry and why we are different.

2.4 Technology

Technology plays an important role in the financial software industry. Because financial data is highly confidential, security is critical in the industry. The security technology should consist of several layers of protection so that if the product encounters malicious attack or invasion, it's still secure and has time to react. Also, handling streaming data is also important to make sure each transaction is handled correctly, accurately within seconds.

To differentiate from others in the industry, we believe a data driven solution serves us the best. Business today relies heavily on data to make better decisions (Rosenbush and Totty, 2013), and data-driven technology can create value for small business. Therefore, our solution is backed with technology that consists of machine learning, data mining, and data visualization:

2.4.1 Machine Learning

A lot of applications focus on modeling and predicting customer behavior and provide feedback and advice. The key is to choose the best predictor for modeling, and evaluating accuracy of prediction is an important element for the method. QuickBooks stores transaction and accounting data for small business in order to provide more services for them. However, the data small business provides is usually full of missing values. Therefore, it is important to find the underlying pattern behind the data. To tackle the task, we apply machine learning method

that is widely used in industry to continuously provide better services to customers by exploring data. Our machine learning applications will focus on both data prediction and optimization.

2.4.2 Data Mining

Data mining is a method to discover underlying patterns behind data, and it relies heavily on statistics. In many aspects, the method is similar to machine learning, but data mining focuses more on discovering associating rules between different variables. However, QuickBooks contains only structured data that is well categorized but hardly provides information for analysis at the first glance. Therefore, we have to put a lot of effort to process, transform, and model the data, and this effort is also the core of data mining. The preparation work we have done in data mining also establishes a solid foundation for later machine learning analysis. We started from generating some data by ourselves, and this adds an intuitive advantage when we perform our analysis. In addition, we invest time in data cleaning, descriptive analysis, and explorative data analysis to ensure we understand the data thoroughly, and we believe these analyses can form a strong, thorough, and detailed understanding in order to create a useful solution for small business.

2.4.3 Data Visualization

Human eye is better than any machine when understanding visualized data, and we hope to a solution that visualizes business data for users. We believe that many small business owners do not have solid business and statistics training; our survey results also reveal that business owners are not interested in exploring math and business figures. Therefore, determining how to

present our solution to users became one of the most challenging tasks to us. To develop the solution, we started from focusing on the easiest and the most intuitive charts, plots, and colors to demonstrate data and its implications. Eventually, we hope to provide interactive and intuitive plotting solution like data driven documents (D3) that help users demonstrate and visualize ideas in easy-to-understand ways.

2.5 Market Definition

The target market of our solution is small restaurant business. According to Statista, one of the world's largest statistics portals, there were around 650,000 restaurants in the USA in spring 2014 (Number of restaurants in the United States from 2011 to 2014, 2014).

The primary stakeholders of our capstone project are Intuit Inc., small business owners, and the capstone project team. The success of the capstone project can be further developed when being integrated into QuickBooks, which is owned by Intuit Inc., and increase Intuit's impact in the financial software service industry. Small business owners will benefit from this product through having a more efficient way of running their business. Members of the capstone project team will be benefited from acquiring knowledge and skills. In addition, QuickBooks' competitors such as Xero and Square are secondary stakeholders as they might be affected by this differentiating positioning of QuickBooks.

2.6 Go to Market Strategy

The definition of go-to-market strategy is to enter the market with a minimum viable product and continuously improve on it. A Minimum Viable Product (MVP) is a product that has just enough features to understand a certain customer behavior (Ries 2011). Providing a minimum viable product will enable the validation of the assumptions that was made on what brings the most value to customers. Moreover, we will be able to use data to enhance the quality of our analysis. Thus, through validated learning we will continuously enhance our product to meet the demand of our customers. This reflects on the 4Ps of marketing mix, which are product, place, promotion, and price.

The second dimension of the 4Ps is the place. Since we envision our product to be an add-in on Intuit's QuickBooks, customers will be able to find us within QuickBooks. Thus, our distribution channels will be integrated within QuickBooks distribution channels, which are online and direct sales. In terms of promotion, we are going to divide our customers into two main groups, existing users and new users. The best way to reach existing users is through direct promotions. This can be attained through e-mails, phone calls, text massages, or advertisement on their QuickBooks account. On the other hand, we will target new customers through online targeted advertisement. By targeted advertisement, we mean customers who already have small restaurant businesses or who are planning to start a new small business. Target customers can be reached through the websites and pages they visited. For example, if they had navigated Intuit's or another competitor website, YouTube page, Facebook page, or LinkedIn page, then they will definitely be customers that we need to target to. This strategy is especially beneficial since it is

much cheaper than mass advertisement. However, in order to advertise we should have a set pricing strategy.

There are three main pricing strategies that are commonly used in pricing any product, competitive analysis, value-based pricing, and cost-based pricing. Since there are no direct competitors that we are aware of, we will eliminate competitive analysis as a viable pricing strategy for us. In the value-based pricing, the price of the product will be determined by how much value this product brings to the customer. Whereas in the cost-based pricing, the price of the product is determined by how much the product cost to be built plus the profit margin. However, since our product is a software, it makes perfect sense to use value-based pricing as our pricing strategy. The value of our product can be determined from the cost saving that small businesses will have due to the use of our product. However, we should keep in mind that it is an add-in on Intuit's QuickBooks, thus the pricing should be integrated with the pricing strategy of QuickBooks. The existing prices for QuickBooks are, $9.99/month for simple start, $19.99/month for essentials, $29.99/month for QuickBooks Plus. Thus, adding a new category would not be a viable idea. However, our product will fall under the QuickBooks Plus category where it will add so much value for customers. We believe that because of our product, some customers will be attracted to the Plus category, and thus generate extra revenues to Intuit.

2.7 The Power of Buyers and Suppliers

The bargaining power of buyers is considered to be weak since this product is solving one of their main pain points and cutting some of their labor costs, and the available solutions in

the market are sparse. According to C. Enz's research (Enz, 2004), 30 percent of people running the restaurant have concerns on labor cost, the highest rank among accounting concerns. She states that "Labor costs are the key issues for everyone." (Enz, 2004). In addition, many solutions in market do not provide analytic functions on efficient operations. Moreover, as the price for QuickBooks Plus is only $29.99 per month, which is equivalent to around 2 hours of labor saving, a high demand is expected to be easily generated. Since such solution is needed, a high demand is easily generated, and the choices are few, the bargaining power of buyers is considered to be weak. To further weaken the power of buyers, we will have to establish a competitive advantage that is unique, and that competitors are not able to imitate. We will provide our premium position by developing highly analytic solution and position it as a financial big data of QuickBooks. Besides, since we will develop an optimized staffing system that connects with employees' paycheck accounts, it will enhance switching cost for customer and thus lower buyer's power; this will not only provide an incredibly designed solution for staffing but also make it much more difficult to switch to other solutions. To sum up, our strategy to lower the power of buyers is to strengthen our competitive advantage and increase their switch costs.

The power of suppliers is also considered as weak for our product. Our solution is data-driven software application, and the suppliers are the users who have data input. However, to use our solution, users must provide data, and as a market leader, QuickBooks already has the big data. Therefore, our strategy is continuing customer acquisition by free trials, friend recommendations, and bundle services with other QuickBooks applications. The more customers

we have, the stronger the network effect, and the weaker the power of our suppliers who are the individual users.

2.8 Market Trend

One of the major trends is the increasing use of mobile devices. In the past, desktop was the primary device in small business operations, but the emerging devices such as smartphones and tablet are changing the landscape. Therefore, in addition to developing an application for desktop, mobile applications must also be developed to meet the trend. Moreover, to better manage diverse devices and to integrate with existing networks, managing the cloud is also needed. According to Victor R. Garza's article in the PCWorld "Mobile devices for business are a dynamically changing market. Eventually, many companies will manage their tablets and phones in the cloud." (Garza, 2011). As a result, the impact from cloud management cannot be ignored. However, QuickBooks' information technology and cloud service are provided by external service providers. Therefore, we must also be aware of the risks of data leakage and server failures.

2.9 Trend Leveraging and Risk Mitigation

To leverage the trend of increasing usage of mobile devices, we will make a responsive web prototype, which means that the website layout can adjust with the size of the mobile device, thus creating a mobile-friendly user interface. We will also try to turn it into a mobile application in the future. With the mobile application, small restaurant owners can view and manage the staffing anywhere anytime. The benefit from this trend is huge. The dynamic staffing

works best if the restaurant owner can check prediction results frequently, which is much easier to be realized on mobile devices, as it is accessible anytime with just a click (Hamerman, 2010). We can add features such as automatic reminder to simplify the business management process. The convenience provided by mobile devices can motivate our clients to use our product. We will also make use of cloud computing in the long term so that we can know the feedback from users and their preference in real time and take actions to maintain the customers. In terms of mitigating the risk of business interruption, we believe that this is a long-term effort dependent on Intuit Inc. Intuit should take actions to strengthen the business partnership and start to build their own devices.

2.10 Effect on Industry Structure

Our product will impact the small business financial software industry and has high potential to reshape the small restaurant industry as well. Originally, the QuickBooks is a pure accounting product which records the transaction data. Our product expands the functionality of the QuickBooks by providing labor management suggestions. Once our product succeed, the QuickBooks can leverage the opportunity to change into a business management software that takes care of all aspects of the small restaurant daily operation. According to Porter's article, one of the five best types of market positioning is to "serve all the needs of a sub-group of customers" (Porter, 1996). By turning the accounting software for into a well-rounded software package that serves all the needs of a sub-group of small business, the QuickBooks will gain great competing power. To compete with QuickBooks, the competitors are expected to follow

the trend and integrate more features into their accounting software. As a result, the barrier to new entries to the financial software industry will become even higher.

The potential impact on the restaurant industry is also anticipated to be significant. By using our product, restaurant owners can hire part-time employees flexibly according to the highly optimized schedule that match the labor supply with customer demand perfectly. On one hand, restaurant owners can save their time and effort on staffing management. On the other hand, restaurants can retain more customers, as our optimization model avoids long waiting time. As a result, the income of the restaurants will increase.

3 Intellectual Property Protection Strategy

Filing a patent is not an effective way to protect our project's intellectual property. We are doing a data analysis project in software, but it is easy to get around software patent with changes in a few lines of codes. In addition, filing a patent is not a conventional way to protect software intellectual property. In 2008 Berkeley Entrepreneurship Survey 400 entrepreneurs, only 30% of the participants from software industry thought that filing patent was an effective way to protect the intellectual property (Graham, Merges, Samuelson, and Sichelman, 2009).

There are two other ways to protect our intellectual property, using copyrights and keeping trade secrets. Each method has its own advantages and disadvantages. The main advantage of using copyrights is that it automatically applies to works that are original and fixed in an arrangement of expression. Therefore, in contrary to patents, registration is not mandatory for copyrights. However, the disadvantage is that it only protects the expression of ideas but not

ideas themselves. On the other hand, trade secret can better protect our outcome if we are able to protect our project's secret sauce: the data and the algorithm. An important feature of protecting trade secret is that we do not need to disclose our secret sauce as we have to when filing a patent, and this is the major reason why some companies do not file patents because confidential information such as solution formula or technology is required to disclose. Competitors could hence learn the technology immediately from the information and develop competitive advantages in the market. Moreover, in contrary to patents, trade secret does not expire. A good example of trade secret that has existed for decades is Coca-Cola recipe. However, keeping a trade secret will hinder the opportunity of licensing the technology to other companies. Furthermore, if the trade secret leaks, the company would lose its potentially edge position and competitive advantage in market. Nevertheless, since we are not looking for licensing opportunities, the best way to protect our technology for the project is to keep our trade secret, our data and algorithm.

3.1 The Risks of the Protection Strategies

There are two risks associated with keeping our trade secrets, the data and the algorithm. First of all, the data might be leaked. Intuit Inc. outsources some of its data solutions, hence the data security is dependent on its business partners. If the data were leaked, the consequences would be severe. If the data that we spent huge amount of time to collect were accessed by our competitors, they could develop the same or an improved product to compete with us and gain competitive advantage in market. More importantly, we would lose our users' trust. The data

Intuit keeps is important to business success as it records how the clients' businesses are running. Therefore, data leakage is a kind of risk that we cannot afford.

The algorithm and business model of our product might be copied, but the consequence will be less severe. As discussed in our strategy paper, the threat from new entries is low because our core competence comes from the huge customer base and data of QuickBooks. Switching cost to other products is high because of user contract, differences in user experience, and data conversion; this establishes a high barrier for competitor and new entrants to gain market share. Moreover, Intuit has collected data from users for more than a decade, and the data could be constantly used to refine our algorithm for improved experience and dynamic and diverse needs from business owners. Therefore, algorithm leakage is not our first concern because competitors do not have the huge customer base to scale up the product nor huge data to constantly improve the algorithm for dynamic needs. Hence, we will put more effort on avoiding data leakage in our project.

3.2 Risk Management

The risk of data leakage can be categorized into two factors: external and internal factors, and we will manage them in different ways. To tackle the risk from external factor, the data leakage through our business partners or users, continuing upgrade on network security and information systems is necessary. Our data is being exchanged between clouds, third-party developers, and small businesses, and this web is dynamic and therefore can't prevent every unethical act. Hence, our information systems should be continuously upgraded to avoid the risk. Moreover, to tackle the risk from internal factor, the leakage through Intuit employees,

improving business operations is critical. Intuit's employees can access the data, and the employees can use the data for illegal or unethical purposes. Therefore, data access should be strictly controlled. For example, different teams have different data access, and exchanging the data or relative information should be prohibited. Since the information an individual employee has been limited, the risk of data leakage is mitigated.

The risk of algorithm leakage is small because the major value of our solution is the data we have. Even if competitors acquire our algorithm, their solutions will not be as effective as ours because they do not have the big data to constantly refine this solution to meet dynamic needs in market. Moreover, competitors do not have huge customer bases to scale up the service, and it is difficult to acquire new customers because of the barrier from high switching cost. Nevertheless, the risk can still be mitigated if the core of the algorithm is segmented and developed by different teams. Consequently, leakage from single or a few teams will not cause severe problem.

4 Technical Contribution

4.1 Project Overview from a Technical Perspective

We live in a rapidly changing world. Emerging technologies are influencing how people think, act, and live. However, it's really the innovative application that drives exciting technologies to the next level. What I really strive to achieve throughout this program is to experiment with technologies I learn and apply it to practical real world problems.

Due to the rise of the Internet of Things (IoT) (Atzori 2010) and creative data exploration, people are empowered with these new applications to live a better life. Of course, business would also benefit from these technologies. The refining and application of data will drive business in a revolutionary way and help generate new perspective and insight. To make the world a better place, our Capstone team has a vision to build useful technology that's related to data science and IoT. Our project is a great example of combining IoT and data utilization to improve business operation.

Partnering with Intuit, our Capstone team positioned ourselves as a team dedicated to helping small business increase their operation efficiency. After weeks of benchmarking various industries, we choose restaurant owners as our targeted customers because they are closely relate to our daily lives and the industry hadn't changed much since a couple of decades ago. Surprisingly, when we interview over ten restaurant owners, they all suffer the same problem: high labor cost is the biggest component of their operational cost and continues to grow throughout the years. To make matters worse, we discovered throughout the interviews that around 90 percent of the time, small restaurants owners are either over staffing or under staffing because they always hire a constant number of people in the restaurant. They have a surprisingly vague estimate on how much people will visit their restaurant within a certain period of time. To tackle this problem, these small businesses are equipping more and more people counting

devices to better understand the customer traffic in their store. These two observations lead us to an exciting idea: If we can accurately forecast the numbers of customers every hour tailored to different type of restaurants, how could we help these small businesses?

Our goal is to solve the complex staffing problem for restaurants through our product. The technical pipeline consists of several stages. First, we need to collect sufficient amount of training data and store it in a distinct dataframe. Secondly, we need to preprocess the data so that it's in good shape for prediction. This part involves a lot of data exploration including normalizing the data, and throwing away outliers, choosing the best basis for our data. At this phase we need several methods to limit the variance, make up for sparse data, and explore fundamental characteristic of the data. Also, we did some visualization work to see the obvious patterns of this restaurant. In phase three, we take in the processed data and run models of predictions. This phase involves optimization problem in order to choose the best parameters for weights for our models. For instance, some sensors might suffer huge bias when they are blocked by the same person. The counting sensor may either count the same person over and over again, or count that person once but fail to take into account the entire crowd of people behind him. In order to minimize unnecessary errors from the sensor data, we need to try out various models and see which model has optimal performance in prediction. We need to setup a baseline and models to evaluate how well our model performs. Also, we need credible test data to see if our models are overfitting or doing well with in the context of test data, the source of truth.

My technical contributions enable our program to take in data from various data sources and, in the future, different types of sensor data, minimize the errors and bias and then generate accurate predictions. The reason why my contribution is important is because different restaurants are using different people sensing product and sensors. Having a product that has the capability to deal with diverse types of data proves its potential to scale. My work involves data preprocessing, and machine learning predictions. Models I've experimented with include Support Vector Machines (SVM), Linear Regression, Polynomial Regression. My technical contribution generates and evaluates prediction outputs, which become input for other teammates to generate dynamic scheduling models. Our team also explored with other factors that are related to restaurant's operation. For example, results from our survey show the time duration customers are willing to wait at a restaurant. With the customer traffic prediction and maximum waiting time as two important parameters, we built a simulation system to generate a dynamic staffing schedule tailored to every restaurant.

4.1.1 How My Technical Part Relates to the Whole Project

My technical contributions focus on two things. First is to process data from sensors and minimize the input variance. My results enable our software to take in data from various data sources and regularize it before storing it as training data. It makes up for sparse data and throws away outliers. It keeps the data clean and ready for prediction. Second part of my technical contribution is predicting the people traffic at a specific time in a day. My predictions generate a

time series of integers, representing numbers of customers visiting the restaurant according to time, that serve as inputs for my teammates operation model (Huang 2015) and together it generate a dynamic staffing schedule. The prediction of customer numbers is an important feature that collaborates with our team's time waiting estimation (Chen 2015). These are two most essential variables in deciding how many employees to hire at a specific time.

4.2 Literature Overview

There are several big trends happening in the industry nowadays. The cheap production cost for sensors and low computation cost provided by vendors, for instance, Amazon, is making new IoT applications possible. On the other hand, the ultra low operation cost for processing scalable data is allowing technology, which was previously unaffordable for small business, to be applied by traditional or family business. In the literature review part of my technical paper, I will explore several relevant papers and articles that add to the significance of our project. In the end I would sum up these trends and applications and state how we are make distinction from our competitors.

4.2.1 The Internet of Things

The Internet of Things(Atzori 2010) consists of various traditional components, which include sensors, actuators, electronic or mechanical devices in various size, interacting with each other. The significant impact for IoT is that these applications can potentially disrupt how people live and how business operates. Within ten years, nodes in Internet will contain food packages, furniture, paper, document and a lot more. Although these technologies contain risks in privacy and security, the miniature design for new IoT components, usually with low energy consumption and cheap consumption make the deployment of IoT an inevitable trend. The propelling visions behind IoT is composed of three components. In the Thing Oriented Vision, RFID is an atomic component that bridge between digital and real world. For the Semantic Vision plays a key role when it comes to extracting structures and meaning behind information that is exchanged and communication between objects. As for the Internet vision, it's comparatively mature to the other visions. Technologies that underlie these visions include identification, sensing, communication, and middleware technology. As shown in Figure2. , applications will be implemented in fields like transportation, logistic, e-heath, personal and social, Futuristic.

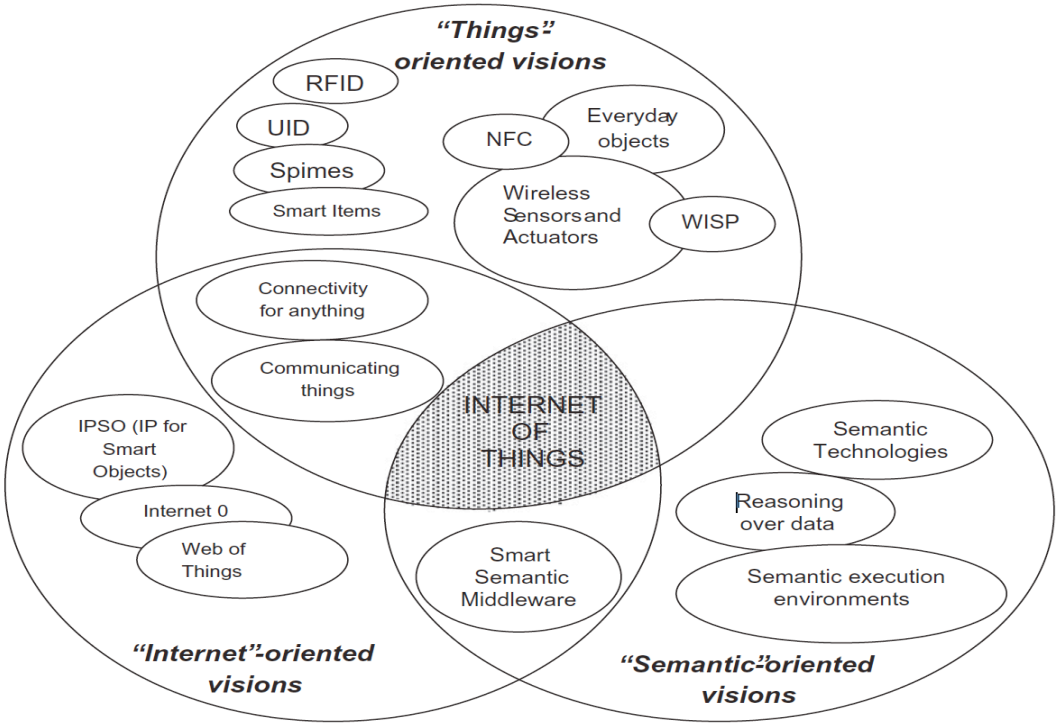


Fig. 1. "Internet of Things" paradigm as a result of the convergence of different visions.

Figure 1."Internet of Things" paradigm as a result of the convergence of different vision

(Atzori 2010)

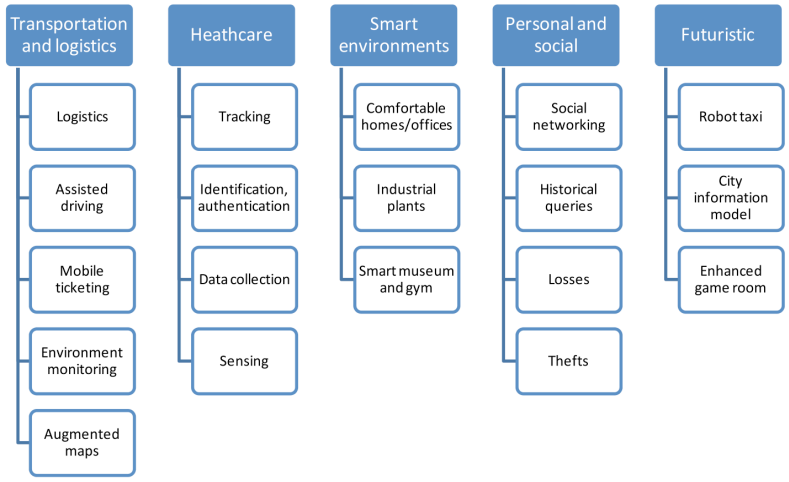


Fig. 3. Applications domains and relevant major scenarios.

***Figure 2. Applications domains and relevant major scenarios*(Atzori 2010)**

4.2.2 Operation Research in Machine Learning and Data Mining

In the past few years, machine learning and data mining is attracting dramatic increase of interest due to the increasing availability and accessibility to data through data collection (Elsevier 2010). Data mining and machine learning is defined as the automation process of extracting previously unknown knowledge and patterns from large dataset. At the core of these methods are some optimization problems, which are also the focus issues in operation research. These optimization problems are a major part of regression, classification, and data clustering.

The intersection between machine learning and operation research lies heavily on optimization problems. Mathematical Programming and Support Vector Machine is used to define an optimal hyperplane. It separates classes of data by finding the maximal distances between data. In other discrete cases, Metaheuristic and combinational optimization is used in complex optimization problems. Usually we obtain an initial solution from the heuristic and utilize different methods to improve the solution.(Ngai 2009)Business can benefit from these approaches starting off by defining the problem they want to solve. After the operation problem is defined, they setup a pipeline to collect data, preprocess data, and then start tackling these optimization problems by applying statistic and mathematical models. Usually this approach will either result in increase in sales or improvement in efficiency.

4.2.3 Application of Operation Research in Staffing and Personnel

The staffing and scheduling problem has been studied for decades due to its major proportion in the overall operation cost. The importance of scheduling has increased these years because company offers more part time contract for flexible work hours. According to Ernst's research on staff scheduling (Ernst 2004), three types of models are explored. They are shift scheduling, day-of scheduling, and tour scheduling. Shift scheduling includes non-overlapping shifts, which is most simple, and overlapping shift. However, due to fluctuating demand over a period shorter

than a shift, over-lapping shifts generate better resource allocation solution. In day-of scheduling, the operation weeks of a business (usually 6-7 days) does not match the working week, which is normally 5 days a week. More assumptions need to be made to constrain this type of problem. Measurement in these models consists of its flexibility in scheduling, constrains need to follow, and performance in terms of cost and revenue.

4.3 Datasets, Methods and Pipeline

4.3.1 Datasets Used(Asha and La Vals)

In this report, we carried out our quantitative and qualitative analyses utilizing two datasets, **Asha** and **La Vals.** The first dataset Asha is comprised of sensors data recorded at Asha teahouse. Here we are able to collect data because the sensor generates a beep sound when a person walks through the door. From our observations we discovered that sensor data is biased either when someone blocked it or when the same person continues to walk in and out of the store. The dataset consist of 390 data points where each data point represents the number of beep sound we recorded within a 15 minutes period of time. The second dataset, recorded at La Vals pizza restaurant, consists of transaction numbers according to a specific timeframe. In consist of 68 rows and 2 columns. One column is transaction number and the other is time. The purpose of this dataset is to predict the people traffic given in a specific timeframe.

4.3.2 Stochastic Gradient Descent

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t).$$

Stochastic gradient descent is an efficient method to solve machine-learning problem, and it has outstanding time complexity performance when applied to large-scale dataset. Stochastic gradient descent is a simplified version of gradient descent because it only takes in a randomly picked data point to calculate its gradient in each iteration. As shown in Equation1, $Q(z_t, w_t)$ is an objective function we aim to minimize. The gradient will change according to your definition of objective. γ_t is obtained by multiplying step size and the value of a randomly selected point. In each iteration, the gradient of the objective function updates the weight w_t until it converges. After calculating the weight iteratively, the optimal weight is tuned with minimized value. In our project, an optimal curve to summarize past sensor inputs is determined using this method.

4.3.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning model in machine learning that performs both classification and regression. SVM constructs a set of hyperplanes to separate the training data into certain categories. After the hyperplane is determined, a label is assigned to a data point based on which side of the hyperplane the data point falls into. Due to the regularized parameter in SVM, one main advantage is to avoid over-fitting. However, in many cases, the entire dataset is not linearly separable. This issue is solved by the kernel method, which is

defined as a mapping function that project data into higher dimensional space. In other words, we introduce non-linearity to our model in order to increase sparsity in a high dimensional basis. This transformation results in a higher accuracy on the test data while increasing computation time.

4.3.4 Setup a Technical Pipeline

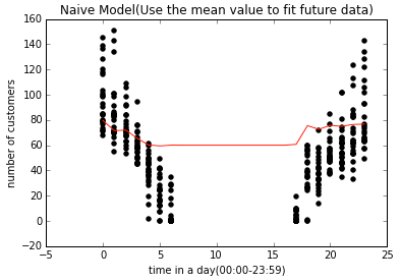
Building a pipeline for my code has several advantages. First, it builds a framework that clarifies every process for our product. Secondly, a pipeline makes separate modules for each process, so it's easy to change models. It outlines a clear structure for our project and is easier to maintain as well as adds new features to it.

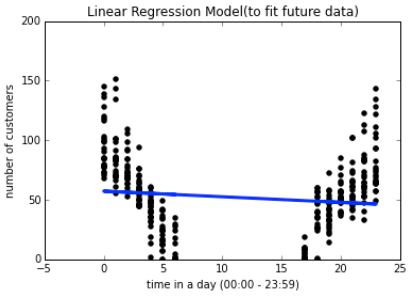
The first stage in our pipeline is data preprocessing. In this stage my code takes in an Excel or CVS file and creates a summary for the file. It determines how many days of data are in this file and tells all the basic statistical characteristic of the file. It shows mean, variance, sparseness, which provide user a basic understanding of the data. Also, in this stage I add some data exploratory features for users to visualize. For example you can easily see the plots of customers coming in and out for one day to compare with another day. It provides most of the basic features needed for clients to get a whole picture of the variation of customer numbers in a day.

In the second stage, which is the most important part, I utilize different optimization methods to minimize the errors of each feature. We do that by building different optimization algorithms, for instance, Batch Gradient Descent, Stochastic Gradient Descent. We evaluate accuracy of our

model with a least square objective function. I built a baseline model for comparison purpose, which I discuss in the following section. After going through this pipeline I built the data so it is readily available as the inputs for dynamic staffing schedule.

4.4 Results and Future Work





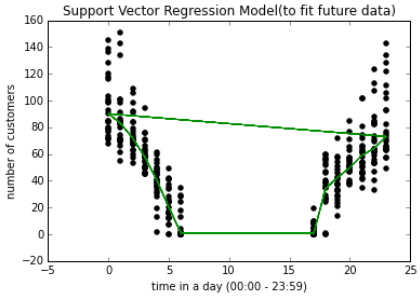


Figure 3. Naïve Model ***Figure 4. Linear Regression Model*** ***Figure 5. Support Vector Regression***

Model Type	*Naïve Model (Figure3.)*	*Linear Regression Model (Figure4.)*	*Support Vector Regression Model (Figure5.)*
Least Square Error	*573.47*	*761.53*	*262.13*
Covariance Score	*0.11*	*-0.06*	*0.42*

Table 1. Performance of each model

4.4.1 Minimize errors from sensor data (Asha)

At this stage, we perform analysis on the first dataset Asha, as mentioned in the methods section. The motivation behind this analysis is to discover an optimal model that removes errors. The errors took place when sensor records data incorrectly. Also, the optimal model should have a minimal least square error after fitting to our data. In other words, we aim to discover a model that gives the best representation of the data. Finally, our software reshapes data into a form as prediction inputs. In real world applications, data is prone to be bias in many situations. For example, if people-counting sensor in a restaurant is blocked by a man standing near the door for a long time, it fails to take into account the people walking pass, and would likely to result in a much lower result.

Figure 3., Figure 4., and **Figure 5.** represent results of the Naïve Model, Linear Regression Model, Support Vector Regression Model. The scattering dots in the graph represent the numbers of customers the sensor detect during each hour. We have three weeks of sensor data of a store and we categorize the data according to hours in a day. As shown in the x-axis, there are 24 categories from 0 to 23, each representing an hour in the day. The scattering dots are the actual data and the curve is the result of the model fitting to the data. The y-axis represents the number of customers visiting your store within 15 minutes.

As shown in **Table 1.**, we calculate the least square errors and the covariance for each model. The least square errors represent the sum of square difference between the prediction value and the actual value. Small least square error implies high prediction accuracy while large least

square error reflect over-fitting results. On the other hand, covariance score reveals the relationship between time and numbers of visiting customer within 15 minutes. Positive covariance means numbers of customer increases along with time. If covariance score is negative, numbers of customer decreases as time increases.

Naïve model is set as a baseline since we simply average our Asha dataset to obtain the result. Least square error is **573.47** for Naïve model and we aim to search for a model that minimize the result. However, as we train with least square model, the least square error increases to **778.74**, and the variance is **-0.27**. This is a proof that linear model has poor prediction performance when dataset is sparse. To improve the performance of our prediction, we also explored non-linear models. As a result, non-linear models, such as Support Vector Regression (SVR) models works relatively well. The least square error of SVR decreases to **262.13**, and the absolute value of the covariance value increases to **0.42**. The result shows that SVR has better prediction performance then the naïve model. Also, time and customer numbers are highly correlated in SVR. As shown in **Fig 5.** ,non-linear model (SVR) gives a better summary of the data because the curve visually fits better to the data.

To sum up, this model will help to filter outlier data and regularize data that's high variance. When a new data point comes in, our program compares it to the value of the past data according to its time, and decides whether it's an accurate or biased data point. For example, if the value of a new data point at 10:30am is 20, and the prediction of our model shows it should be 100 customers at this time, we know the value coming from the sensor is likely to be biased since it's 80% below our prediction. The efforts here will later be helpful for following part when we aim

for accurate machine learning predictions. One problem here is that it's hard to generalize the data processing step when different sensors record data in disparate format. However, my code is flexible enough to deal with various frequencies.

4.2.2. Predict numbers of in-store customer at a specific time (La Val's)

```
x_test = np.array([ 10.25,  10.50, 10.75, 11.0, 11.0, 11.0, 11.25, 11.25, 11.25, 11.5, 11.5, 11.5, 11.5,
       11.75, 11.75, 11.75, 12, 12, 12, 12.25, 12.25, 12.25, 12.5, 12.5, 12.5, 12.5,
       12.75, 12.75, 12.75, 13, 13, 13, 13.25, 13.25, 13.25, 14, 14, 14, 14,
       14.25, 14.25, 14.25, 14.5, 14.5, 14.5, 14.75, 14.75, 14.75, 14.75, 14.75, 16.5, 16.5,
       16.75, 16.75, 17, 17, 17, 17.25, 17.25, 17.25, 17.5, 17.5, 17.75, 17.75, 17.75,
       18, 18, 18]).reshape(1,68)

y_test = np.array([ 40,  75, 128, 140, 134, 143, 145, 156, 163, 163, 166, 169, 171,
       176, 177, 180, 183, 185, 185, 188, 190, 192, 194, 197, 200, 203,
       204, 206, 213, 214, 216, 250, 220, 224, 225, 228, 231, 231, 232,
       235, 250, 250, 242, 243, 247, 250, 255, 264, 260, 274, 300, 349,
       354, 392, 362, 365, 367, 370, 378, 398, 460, 404, 406, 408, 421,
       424, 496, 516]).reshape(1, 68)
```

Figure 6. x_test is the prediction input , y_test is the prediction output

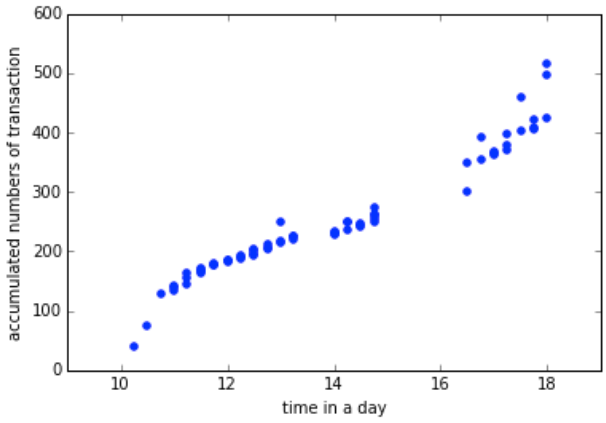


Figure 7. Support Vector Regression Model for Prediction

In this section, my goal is to predict the accumulated numbers of in-store customer at a specific time in a day. We carried out prediction using the La Val's dataset. As shown in **Figure 6.,** the data consist of two weeks of La Val's transaction data from 10am to 6pm. Note that the transaction number is accumulative throughout the day. For prediction purpose, I split the data into training data and testing data. Training data fits time to the numbers of customer according to our Support Vector Regression model, the optimal model we selected in the previous section. Test data serves as a source of truth to evaluate, after training the model, how accurate the prediction output. As shown in **Figure 7**., the prediction values of transaction number increase in regards to time. The average value of the prediction value is **251.132**, while the mean difference between prediction and labels is **4.22**. This implies that given a specific time in a day, our model could accurately predict the numbers of customer in a store with a **1.68%** of error rate. Utilizing the good prediction result alone, restaurant owner would have a much better sense of how many employees to hire for a specific shift. Furthermore, combined with the waiting time analysis in our project, these results would provide much more insights for our customers.

For future endeavor, defining other interesting features would empower interesting predictions. For instance, if you have all the table numbers or geo-location data, you could possibly map our prediction result on the map and suggest the nearest restaurant that has highest probability of lowest waiting time.

5 Conclusion & Reflection

I learnt that dealing with real world data is a huge challenge. In many real world scenarios the data collected by sensors are noisy and sparse. Therefore the preprocessing stage is essential in terms of throwing away outliers, selecting tradeoff between bias and variance, and normalization uncharacteristic data. Another lesson I learnt was that having the data is essential for any software service. The shortage of data would not only lower your prediction accuracy but would also limit your creative endeavor. Sparse data wouldn't be useful at all for models like data clustering to regression. The following question is: How would you create extra value from data? It turns out that defining feature is much harder than we expected. In real world situations you define all the variables on your own. Redundant features lead to noise and misleading result, while lack of features could result in meaningless result. After doing the technical part of our project, I got much more familiar with all the tools and became better in modeling real world situations to machine learning problems. To sum up, these creative processes helped me improve my ability to understand customer needs and tackling pain points for client. More importantly, working on technical parts with teammates in different discipline increased my communication skill and leadership.

The project outcomes are very different compare to our original plans. These adjustments made throughout the year made our project more practical and customer focused. First transition was the change of advisor that took place last semester and the other issue was the lack of data we receive from Intuit. However, I am proud to say that, although our final results distinct from our original plan, it's a positive thing because we learnt and grown to adapt to changes we made and

became much more realistic and practical dealing with our capstone project. We used a lot of design thinking methodology to test our ideas. We have brainstorm meeting every week and verify our ideas with restaurant owners, whom we regard as our potential customer. The upside of this methodology was that we came up with various creative ideas, while the downside is the constantly changes in our plan create extra workload for everyone. Overall, we created a deep sense of customer empathy, and moreover, become much more capable of implementing the technical aspect of our project.

For future endeavor, I would suggest a more well-define topic collaborating with the industry might be a better. If that's the case, we are allow to focus more on the technical parts and apply what we learn to our project. If someone were to pick up our project, I would suggest him or her to start from the architecture we designed and tried to implement it by building prototypes.

In conclusion, this was my first one-year long project and I honesty learnt so much from the start to the end. Especially, I would like to thank my teammate for being open-minded and collaborative. My Capstone experience made me a better engineer and communicator as well. Therefore I am full of gratitude to Fung Institute, and people who've helped us along the way.

References

Akyildiz, I. F.

"Wireless sensor networks: a survey." Computer networks38.4 (2002): 393-422.

Atzori, Luigi,

"The internet of things: A survey." Computer networks 54.15 (2010): 2787-2805.

Barrett M. P.

2014 The Cybersecurity Myths That Small Companies Still Believe. http://www.bloomberg.com/bw/articles/2014-11-24/the-cyber-security-myths-that-small-companies-still-believe, accessed March 1, 2015.

Bierlaire, M.

2009. Estimation of discrete choice models with BIOGEME 1.8. Ecole Polytechnique Federale de Lausanne, Transport and Mobility Laboratory.

Brown, T.

2008. Design Thinking. Harvard Business Review

Chen, X.

2015 Disovering Insights from Data Analysis.

Edwards, J.

2014 IBIS World Industry Report 54121b: Payroll & Bookkeeping Services in the US. http://www.ibis.com, accessed February 28, 2015.

Edwards, J.

2015 IBIS World Industry Report 54121c: Accounting Services in the US. http://www.ibis.com, accessed February 28, 2015

Enz, C. A.

2004 Issues of Concern for Restaurant Owners and Managers. Cornell Hotel and Restaurant Administration Quarterly, 45(4), 315-332. http://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1360&context=articles, accessed Feb 8, 2015

Ernst, A. T.

"Staff scheduling and rostering: A review of applications, methods and models." European journal of operational research153.1 (2004): 3-27.

Garza, C. A.

2011 Mobile Device Management for Small Business. http://www.pcworld.com/article/240768/mobile_device_management_for_small_business.html?page=2, accessed February 16, 2015Graham, J. H. S. Merges, P. R. Samuelson, P. Sichelman, T.

2009 High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey. http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=3124&context=facpubs, accessed February 28, 2015.

Hamerman, P. D.

2010 Mobile applications will empower enterprise business processes. https://www.forrester.com/Mobile+Applications+Will+Empower+Enterprise+Business+Processes/fulltext/-/E-RES57563, accessed March 10, 2015

Higginbotham, S

2013 Consumers Could Pay Higher Switching Costs in a Data-Driven World. http://www.bloomberg.com/bw/articles/2013-04-12/consumers-could-pay-higher-switching-costs-in-a-data-driven-world, accessed March 1, 2015.

Ho, A.

2015 Discovering Insights from Data Analysis.

Huang, I.

2015 Discovering Insights from Data Analysis.

Intuit Inc. Annual Report for the fiscal year ended July 31 2013.

2013 http://investors.intuit.com/files/Intuit%20FY13%20Form%2010-K%20r221%20at%2009-13-13%20FINAL%20CLEAN%20to%20RRD.pdf, accessed February 16, 2015

Ma, F.

2015 Discovering Insights from Data Analysis.

Moshe Ben-Akiva, D. M.-S.

1999. Extended Framework for Modeling Choice Behavior. *SpringerLink*.

National Restaurant Association

2014. Big Data and Restaurants: Something to Chew On http://www.restaurant.org/Downloads/PDFs/BigData. Accessed on 4/28/2015 at 4:55PM

Ngai, E. W. T

"Application of data mining techniques in customer relationship management: A literature review and classification." *Expert systems with applications* 36.2 (2009): 2592-2602.

Number of restaurants in the United States from 2011 to 2014.

2014 http://www.statista.com/statistics/244616/number-of-qsr-fsr-chain-independent-restaurants-in-the-us/, accessed March 11, 2015

Porter, M. E.

1996 What is Strategy? Harvard Business Review

Ries, E.

2011 The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to. New York: Crown Business.

Rosenbush, S., Totty, M.

2013 How Big Data Is Changing the Whole Equation for Business. http://www.wsj.com/articles/SB10001424127887324178904578340071261396666, accessed February 28, 2015.

Sensource official website

Learning about our Sensors and Data Collection Technology. http://www.sensourceinc.com/technology.htm, accessed February 28, 2015.

Sigurdur Olafsson, X. L.

2008. Operations research and data mining. *Elsevier*.

Smola, A. J.

"A tutorial on support vector regression." *Statistics and computing* 14.3 (2004): 199-222.

Train, K. E.

2009. *Discrete Choice Methods with Simulation.* New York: Cambridge University Press.

Van den, J.

"Personnel scheduling: A literature review." European Journal of Operational Research 226.3 (2013): 367-385.