FINAL REPORT

MIS584 Group 2

FADI ALMAZYDA
YANG LI
MINKUN LIU
YUCHEN SHEN
JINLIAN LIN

APRIL 27, 2018

WORCESTER POLYTECHNIC INSTITUTE

Contents

Section 1: Executive Summary

The goal of this project is to find a client company who could use a Business Intelligence system to improve their decision-making process at the management level and also improve their tactical strategic management process. After acquiring the proper datasets, we built two dashboards and two predictive analysis models based on our research on the client's industry, current operation and BI status. The goal of these dashboards and predictive analysis is to integrate and analyze relevant and useful business information, generate key performance indicators (KPIs) to detect significant events, identify or monitor business trends, so the client can use them to quickly adapt to a changing environment or senario. Another goal we had in mind when building the dashboards are to present the data analysis results in an intuitive, easy to use, interactive and visual appeal way.

We have encountered several problems in the process of developing the project. Firstly, our client chose to not reveal their real business data which is understandable, we had to find relevant datasets online that can be applied to the client's business. After finding the proper datasets, we had to clean it to make sure it fits the client's market. We were also not sure if the datasets we found online can generate enough significant KPIs to give the client useful insights, so we consulted our instructor Dr. Nima Kordzadeh regarding the issue. Secondly, since we are graduate students who have very limited knowledge about our client's industry, we had to do a lot of researches to understand what kind of KPIs their industry typically uses, why these KPIs are important to their business and how to calculate the KPIs using the datasets we have on hand. Thirdly, after doing the predictive analysis on the datasets, we found the results were not ideal. We were able to make predictions but are lack of accuracy. We then had some discussions around what attributes should be used in the model and if there is other predictive model/algorithm that can work better. Finally, after we had our first prototype of the dashboards and showed them to Dr. Kordzadeh, he pointed out that one of the databoards doesn't quite fit into our client's organizational structure. We then redesigned that dashboard and replaced some of the KPIs. Overall, we had a few issues when doing the project, but we were also able to resolve them in a timely manner and produced ideal results.

To summarize the outputs of the project, we built two dashboards, one is tactical and the other one is strategic. Each dashboard has four KPIs that are interconnected to each other, this means users can make a selection on one KPI Figure and view the corresponding results get updated on the other three KPIs Figures. We added several filters to the dashboards to give the user more flexibility to view the data for specific purposes. There are some single data boxes on the dashboard as well and they can give users what the business status is at a glance. We also developed two predictive analysis models that make predictions on two different aspects of our client's business to help them stay ahead of competitors.

Section 2: Introduction

2.1 Client company background

We choose the Jarboe Group as our client company, which is a residential real estate broker located in Massachusetts. They have been in the real estate industry for 9 years now. They provide services for buyers, sellers and investors by helping customers sell and buy houses as residences or as investments. The real estate property they deal with includes both new construction and resale homes.

The Jarboe Group has an office in Worcester, the location of their customers spread mainly throughout Massachusetts and Connecticut. So, the real estate market they serve is mainly around the Worcester area, such as Auburn, Boylston, Northborough and etc.

The Jarboe Group is a small compact team, with the owner, Justin Jarboe and only 6 employees. This includes 4 real estate agents, Justin's wife Alicia, and a dedicated data analyst Alain. Alicia is in charge of the day to day operations of the company, and Alain is a data analyst who used to be a hedge fund manager on Wall Street. By looking at the structure of this streamlined team, it's easy to tell how important data analysis is for the company. Business intelligence (BI) and data analysis are important to the real estate industry as well. Across any company, it gives managers a lot of insight about the business, like the peak time for buying or selling houses, the trend of the real estate price in a certain area, and so on. We chose the Jarboe Group as our client company because of their focus on data. Considering they have a simple organizational structure, the scope of this project is to design and develop a BI system for the whole company to use.

2.2 Current BI status

According to the information provided by Alain Digon, who is the data expert on the team, the Jarboe Group mainly uses Microsoft Excel to track all of their goals. These goals are determined through techniques such as making pivot tables to analyze data. All of their analysis is conducted in Excel. The Jarboe Group don't use any data warehouses or any seamless integrated information system such as an Enterprise Resource Planning (ERP). In addition, they don't have a separate BI team. Their main purpose of using data is to conduct analysis, monitoring and forecasting for their daily operations. They can also determine market performance compared to other competitors and agent performance.

In order to achieve business value from BI, the key factor is realizing the development of the company. The maturity level of Business Intelligence within the company must match as much as possible to the maturity level of the company itself. Only then will the benefit of the Business Intelligence be highest. A figure depicting the appropriate level of maturity is below.

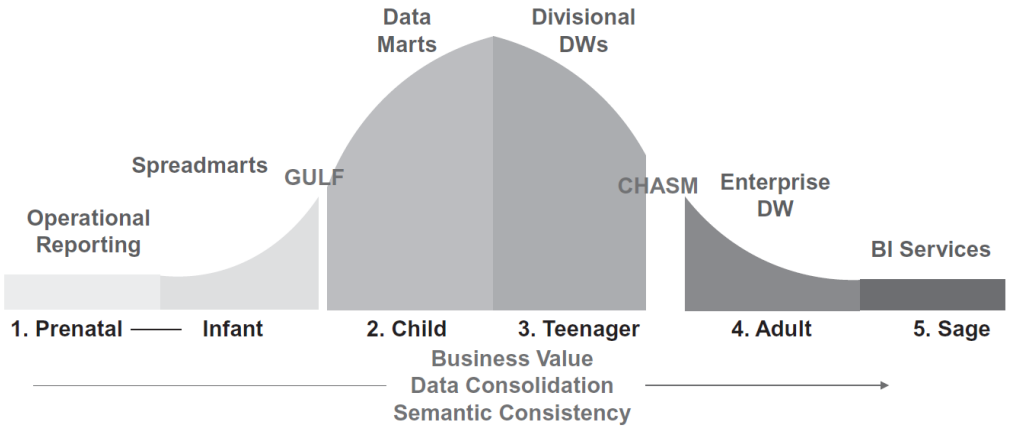


Figure 2.1 BI Maturity Model [1]

As for the Jarboe Group, they are at the first level, which is composed of two phases: Prenatal and Infant. Because the Prenatal phase lasts until a data warehouse is created. Specifically speaking, the Jarboe Group is only in the first level Prenatal phase. No data warehouse is created, their reports usually are built into operational systems and are limited to that individual system.

No data warehouse is appropriate for the current status of the Jarboe Group, because it is a start-up company. They don't have numerous data sources to be processed and analyzed. However, for the long-term development of the Jarboe Group, they will need to create their spreadmarts for dealing with numerous partial data sources? Because fragmented data sources will produce conflicting views on business information, which will undermine the effective decision-making process supported by strategic goals and will prevent a clean and consistent view to all events in the company.

The Jarboe Group's main business is focusing on the Massachusetts area, and little on the Connecticut area. For the future development, they may expand their business into other states in the East or even the whole United States. Combined with their mission and vision, which is to create a trusted name in the community and build the business for funding an appropriate balance of work, life and giving back. To better accomplish their mission and vision, creating data warehouse to collect more housing information and customer feedback is necessary for the Jarboe Group.

Section 3: Proposed Business Intelligence Solution

3.1 Current Issues

Real estate industry has historically depended on outdated and disparate systems, using spreadsheets and Customer Relationship Management (CRM) tools to house mountains of data. However, those systems don't communicate with each other, which is a huge problem. Jarbo Group is not different from other competitors in the market who still using old systems. Jarbo Group still depends heavily on spreadsheets and pivot tables which causing huge problem for managers and agents alike. Currently managers need to spend countless

hours digging through multiple spreadsheets to answer one simple question such as which agent contribute the most revenue and through what type of property did it happen? To answer this question, the manager need to use several spreadsheets to run some data work. This can make it near impossible for manager to see the big picture and make the right decision at the right time using the right information.

3.2 Business Intelligence

To solve this problem for our client, we propose implementing a Business Intelligence tool that help our client to make insightful decisions. "BI encompasses a wide variety of tools, applications and methodologies that enable organizations to collect data from internal systems and external sources; prepare it for analysis; develop and run queries against that data; and create reports, dashboards and data visualizations to make the analytical results available to corporate decision-makers, as well as operational workers"(Tech Target). BI will give our client competitive advantage by providing the right information to make the right decisions. BI tool will give managers access to a high-level view of their entire portfolio from a data and performance perspective. BI aims to add a step of automation in the process of analyzing and aggregating data, helping managers reduce downtime, close deals faster and increase company and agent's performance. Since the industry already enjoys access to massive sources of data, comparable sales, portfolio metrics from Multiple Listing Service (MLS), Redfin and Zillow, we proposed to our client to implement a BI tool from Tableau and feed the BI solution data from these available sources. However, before loading the data to the BI, the data need to be cleaned and preprocessed to make sure it contains accurate data which ultimately will produce accurate information.

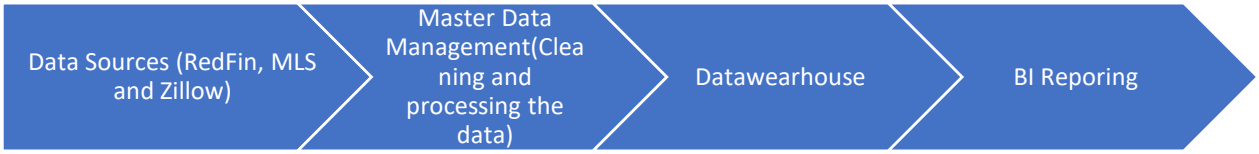


Figure 3.1 Data Flow Process

3.3 BI Components

The proposed BI solution contains eight components which we call them Key Performance Indicators (KPIs). These will be explained in detail later. Before we introduce these components, we will display a star schema that contains all the data that needed to feed the BI solution.

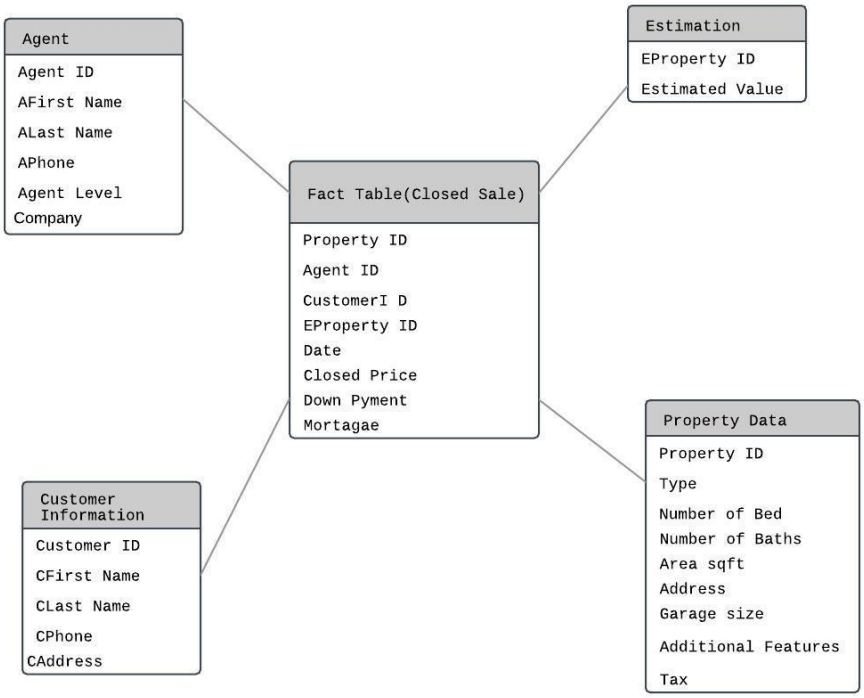


Figure 3.2 Star Scheme

3.4 Dashboards

The proposed BI implementation will include a Market Analysis Dashboard and Agent Performance Dashboard. The Market Analysis Dashboard can help managers to see the big picture of the market. The first KPI for this dashboard is the Sales performance by competitors. The sales are calculated by the closed price times %6 times number of houses sold by each company. These are public data which are available by MLS and Redfin datacenter. This KPI can help Jarboe Group to see where they stand among their competitors in overall sales or over a specific period of time. The Second KPI in the Market Analysis Dashboard is Revenue analysis. This KPI help Jarboe Group to see how their revenue is compared to the market overall and compare it with a specific competitor. Incorporating the time help Jarboe to see their revenue over a specific period of time compared with a specific competitor. The third KPI is Average Sales Cycle by Competitors. This KPI helps Jarboe to see where they stand when it comes to how fast the company can sell a property in average. For example, currently Jarboe in average need 142 to sell a property. Jarboe position currently is fifth in the market among its competitors. The fourth KPI in the Marketing Analysis Dashboard is the Accuracy of Estimating Property Value. This KPI helps Jarboe managers to compare the company accuracy against its competitors. It is computed by the absolute value of the closed price minus the estimated price divided by the closed price. By measuring this KPI, Jarboe managers can assess their agents' knowledge in the market. If their assessment tends to be far from the closed price, then they need to work in improving their estimation technique and methods. Having inaccurate estimation consistently can impact Jarboe group as major competitor in the market.

Customers both sellers and buyers want as much as they could of an accurate pricing for their properties.

The Market Analysis Dashboard provides Jarboe managers with a big picture of the market. It incorporates the time between all the KPIs in it. One more feature in the Market Analysis Dashboard is the interactivity between all four KPIs. A manager can click on one of the competitors and then all measurements in all 4 KPIs run against Jarboe's KPIs. Also, four boxes on the top of the dashboard display summary of any competitor that the manager click on. The data about competitors are public data that is available on MLS.

The second dashboard is Agent Performance Analysis. This dashboard can help managers to assess and monitor agents' performance on a giving period of time. There are four KPIs for this dashboard. First one is Agent performance which display how much revenue is generated by each agent. To dig deeper into this KPI, we connected it with another KPI called Average Sales Cycle by agent. This KPI displays how many properties are sold by each agent and how long does it take each agent to sell a property. These two KPIs can differentiate between top agents' performers and lower agent performers against the company benchmark. These KPIs also help managers to tailor and target training to lower performed agents to help them raise to the expectation of the company. The third KPI is displaying the average price per sqft over time. This will help managers to measure agent performance according to the market status. For example, it is not fair to say one agent sold more than 20 properties but the revenue that this agent generated is in the fifth or sixth on the company ladder, whereas another agent sold 5 properties, but his or her revenue is the second or third on the company ladder. After digging more into the average sqft price, we found that the first agent sold more properties when the market is not performing well as a whole whereas the second one sold all properties during healthy market. The fourth KPI is to display the locations that each agent did business in. This help managers to redistribute agents according to their performance on certain areas. By focusing agents in their best performed area, agents can build more expertise in the market of that specific area or region.

Section 4: Three Use Cases/Prototypes

Our prototypes are created based on a dataset released in Kaggle [2], since the Jarboe Group didn't provide any real data. In order to remain useful details and remove missing value, a preprocessing step has been done based on background knowledge. We convert all address information into coordinate to find relationship with locations. After cleaning and transforming, the new dataset contains 2788 selling records with 36 features about property and selling information, such as price, size and days on market. All these information is usual for a real estate company to collect, and using our dashboard and analysis generated from these data will definitely benefit the company.

4.1 Strategic Dashboard

In order to have a clear view of the whole real estate industry and know the position of your company, we recommend the Jarboe Group to have a BI Strategic Dashboard which can be similar to our prototype shown below. (Figure 4.1.1)

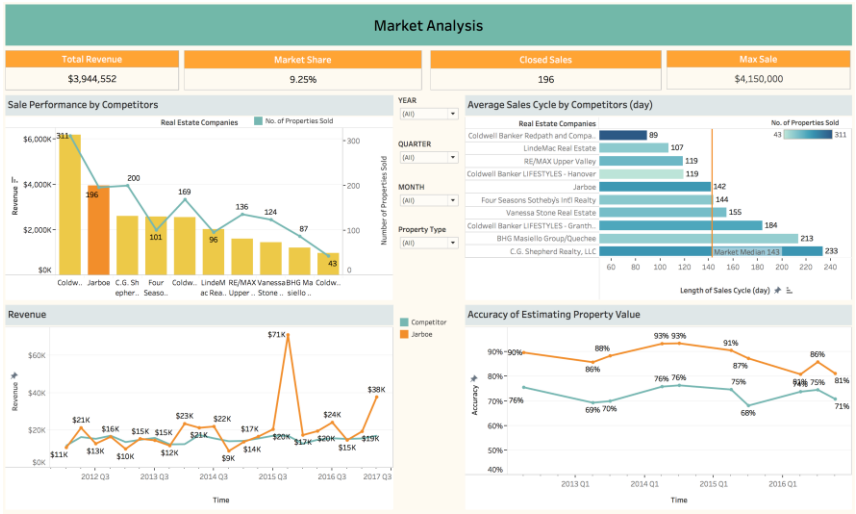


Figure 4.1.1

In this dashboard, there are mainly 5 parts for market analysis and 2 filter parts.
The filter part in the middle gives manager choices to see different aspects of the company. The performance from different types of the property and different time intervals, such as years, quarters and months, can be analysis separately using the filter. This changed dashboard gives the performance in 2016. (Figure 4.1.2)

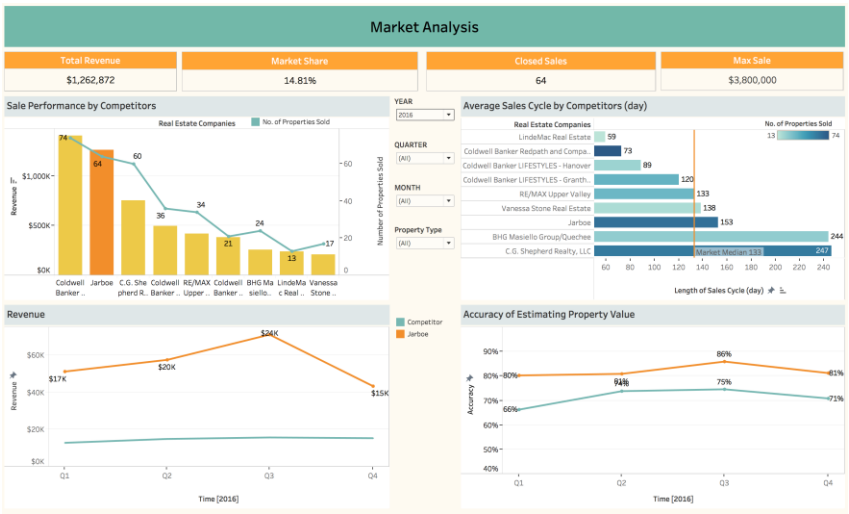


Figure 4.1.2

The other filter is hidden in the first Sale Performance by Competitor sheet. Linking all the sheets, managers can select one of their competitor and compare the performance between Jarboe Group and the competitor. The competitor's performance in Average Sales Cycle sheet has been highlighted, and differently in the bottom two sheet, Jarboe Group's performance is shown in orange and competitor is shown in blue. We design this to make people easier to interpret the

difference and know whether company goes well comparing to the whole industry or competitors. (Figure 4.1.3)

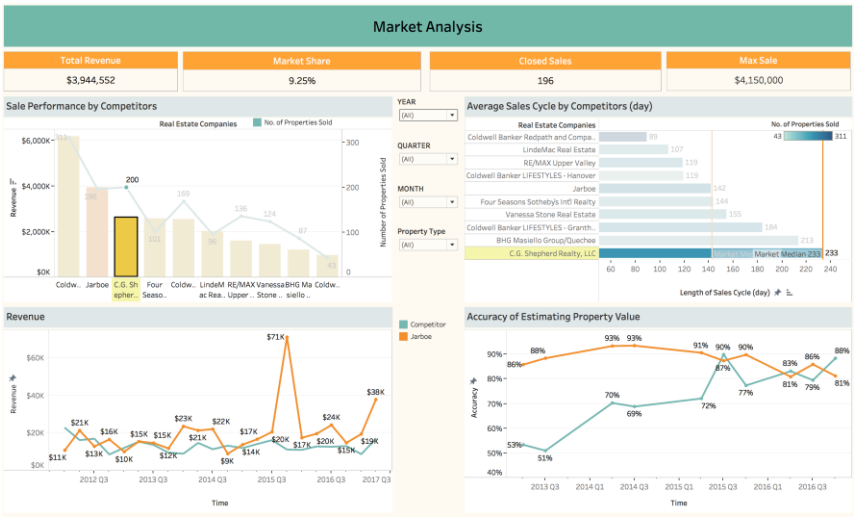
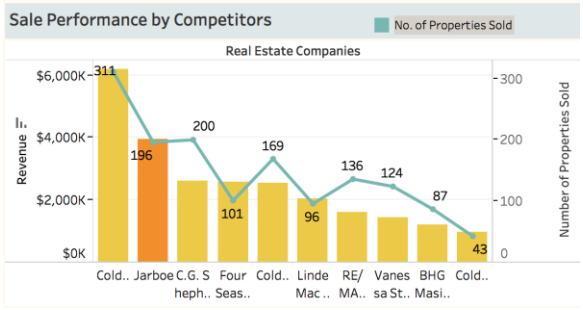


Figure 4.1.3

At the top, there are 4 boxes. Individually, Total Revenue shows the total revenue gained; Market Share is the percentage of the Jarboe Group's revenue in industry revenue; Closed Sales shows how many properties have been sold; Max Sale is the highest final sold price for a single property. All these boxes will change along with the filter and also with the competitor we choose from Sale Performance by Competitor sheet. Clear numbers are important for manager to understand performance. (Figure 4.1.4)

Figure 4.1.4

(Figure 4.1.5) Sale Performance by Competitor sheet is design to be the filter of competitor and also the most important part to show the industry performance. The bar chart is all real estate company's revenue and the blue line is used to give the number of properties sold. A combination of these two metrics can show your position in the industry and give an overview of the type of your company. As an example, RE/MAX Upper Valley company have sold many properties, but low revenue comparatively, means they focus on selling more cheap houses or may be bad selling skills.



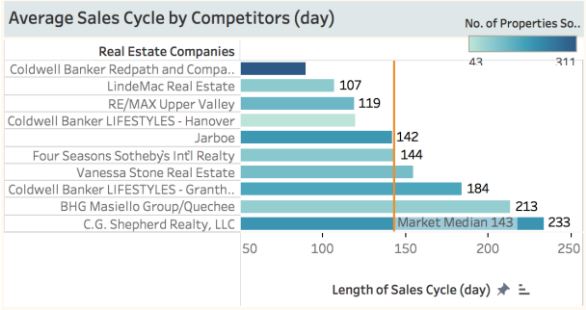


Figure 4.1.5 Figure 4.1.6

(Figure 4.1.6) Average Sales Cycle by Competitors(day) using two metrics, average days needed to sell a property (bar chart) and also the number of properties sold(color). Darker color means more properties sold and shorter bar means selling properties quicker, which is the best combination. Coldwell Banker Redpath and Company has the best performance here and the Jarboe Group works not bad too, lower than the orange vertical line which is the industry median days needed.

(Figure 4.1.7) Revenue sheet shows the revenue along with time, and the orange line gives the entire view of Jarboe group's performance during different time period. Meanwhile, the blue line is the industry average, and when choosing a particular competitor from other sheet, it shows the competitor's revenue instead. Comparing Jarboe Group and industry or competitor with time is useful to know the company growth for future decision.

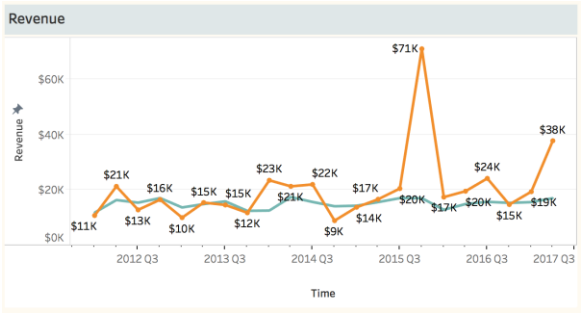


Figure 4.1.7

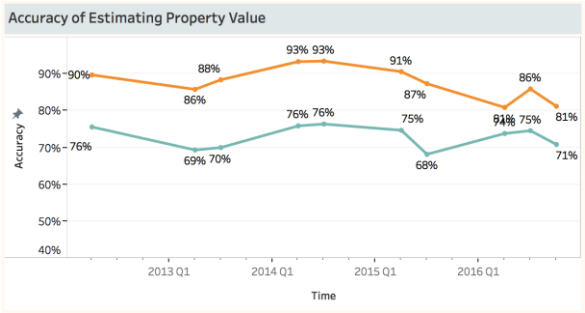


Figure 4.1.8

(Figure 4.1.8) Accuracy of Estimating Property Value sheet is designed by calculating how accuracy between the company assessment price and the final sold price. It's important to predict the price of a property, company will know their potential revenue precisely, and they also can provide the assessment to their customers. Since customer will not be satisfied when having a sold price lower than their expectation which always much higher than real price, an accurate assessment price should lead them to a proper expectation and make them believe in your company. Similarly, with timeline, orange line for the Jarboe Group and blue for industry or competitor.

4.2 Agent Performance Dashboard-Tactical Dashboard

Since the manager can interpret the overview company performance from the strategic dashboard, we recommend the Jarboe Group to have another BI Dashboard which can be similar to our prototype shown below. (Figure 4.2.1) This dashboard is design to show agent performance, agents are the most important part of a real estate company. Managers can easily see the performance of each agent in this dashboard.

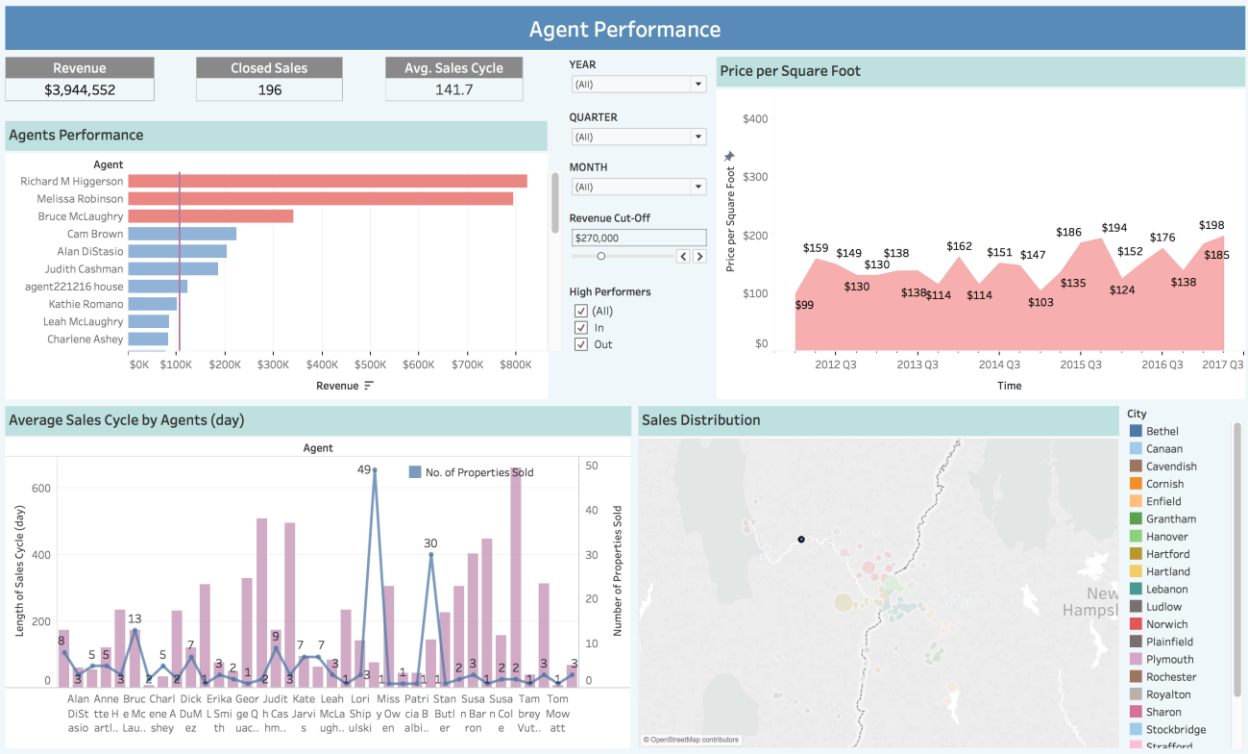


Figure 4.2.1

Similar to the first dashboard, there are mainly 5 parts for agent performance and 2 filter parts in this dashboard.

The filter part in the middle gives manager choices to see agent performance during different time periods. The other filter is hidden in the Agents Performance sheet. Linking all the sheets, managers can select one of their agent and compare the performance between them. The competitor's performance in Average Sales Cycle by Agents sheet has been highlighted, and other two sheets, Price per Square Foot and Sales Distribution, will only show information related to the selected agent. We design this to make manager easier to see which agent works well and which is bad. Helping to make decisions on agent management. (Figure 4.2.2)

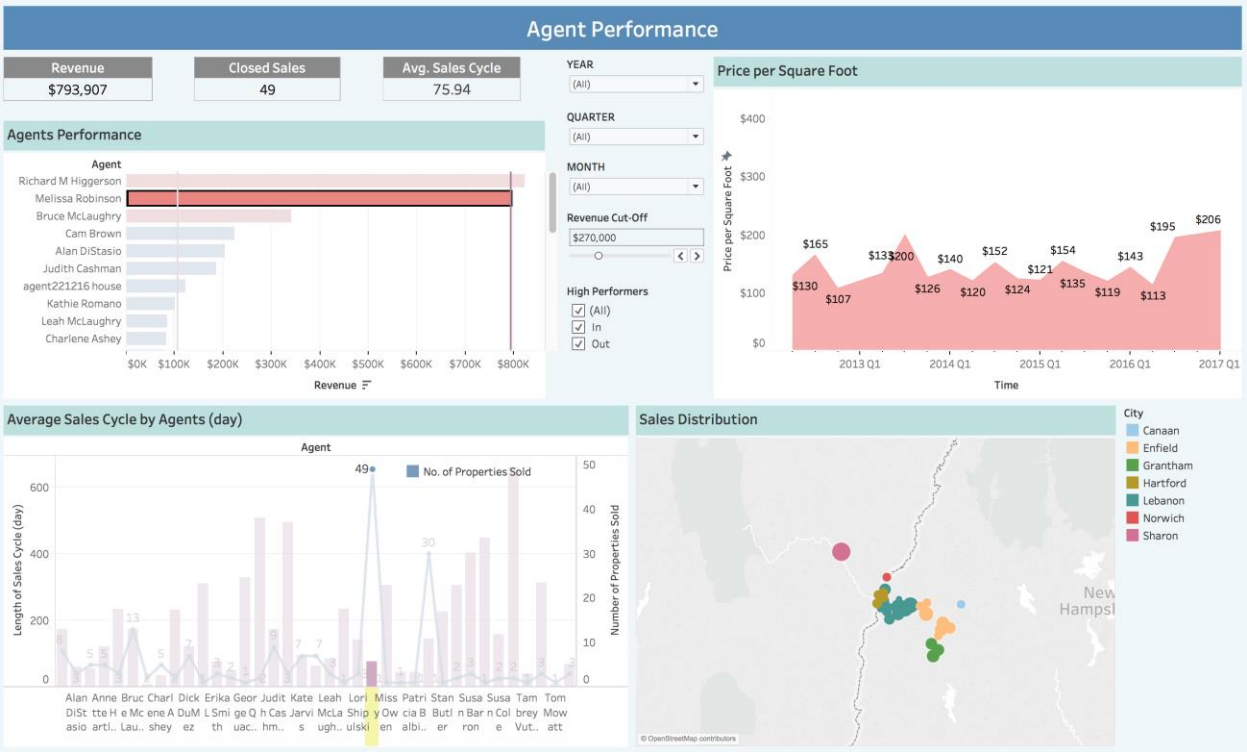


Figure 4.2.2

At the top left, there are 3 boxes related to agent performance. When not choosing any specific agent from the Agents Performance sheet, three boxes will separately show the total revenue, number of house sold, and average days needed to sell the house which generate by all agents. These boxes are related to all filters and will change to give a clear number data for users to understand the agent performance.

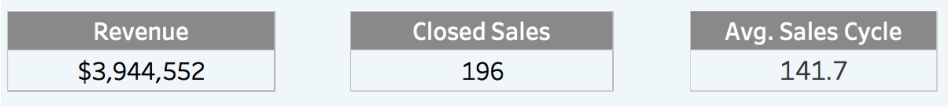


Figure 4.2.3

(Figure 4.2.4) Agents Performance sheet is design to be the filter of agents and a direct metric to rate the performance. The bar chart is the overall revenue generated by each agent and the vertical baseline is average of total revenue by each agent. There are two filters to make agent performance clearer. Managers can use Revenue Cut-Off to set their preferred high-performance baseline, and those high performed agents will be shown in red bar. Moreover, they can choose only to see high or low performance agents using High Performers filter.

Figure 4.2.4

(Figure 4.2.5) Manager will use Price per Square Foot sheet with timeline to see agents selling performance during different time periods. Price per square foot is a proper metric to evaluate the value of agent's properties which is not influenced by the size of house.

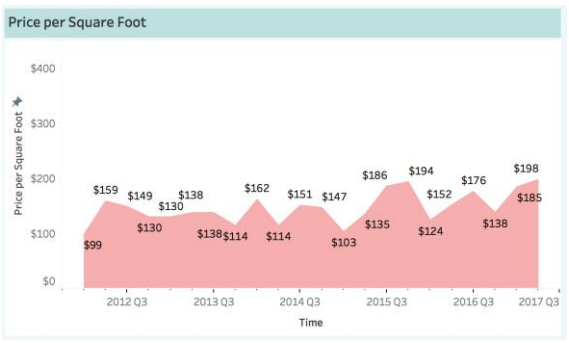


Figure 4.2.5

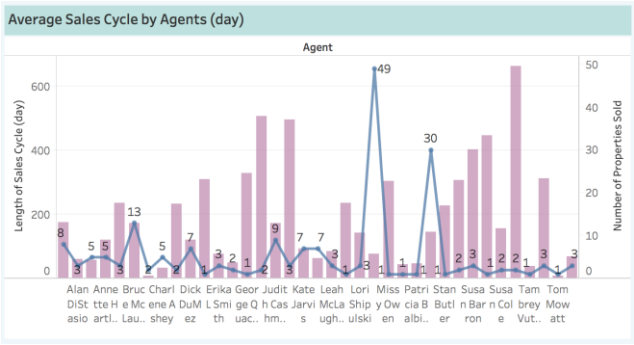


Figure 4.2.6

(Figure 4.2.6) Average Sales Cycle by Agents(day) sheet is similar to the one in Startegic dashboard. The bar chart gives average days needed to sell a property and the line explains the number of properties they sold by each agent. High number of property sold, and low number of days needed gives the best performance.

(Figure 4.2.7) The map of Sales Distribution exhibits the exact location of the Jarboe Group's property. Different colors represent many cities in this area, and when hover on the point in the map, it will give the detail of location along with revenue which can be generate from the property. This Sales is linked to agent filter, clicking at different agents, managers can see where this agent usually sells property and his revenue from each property.

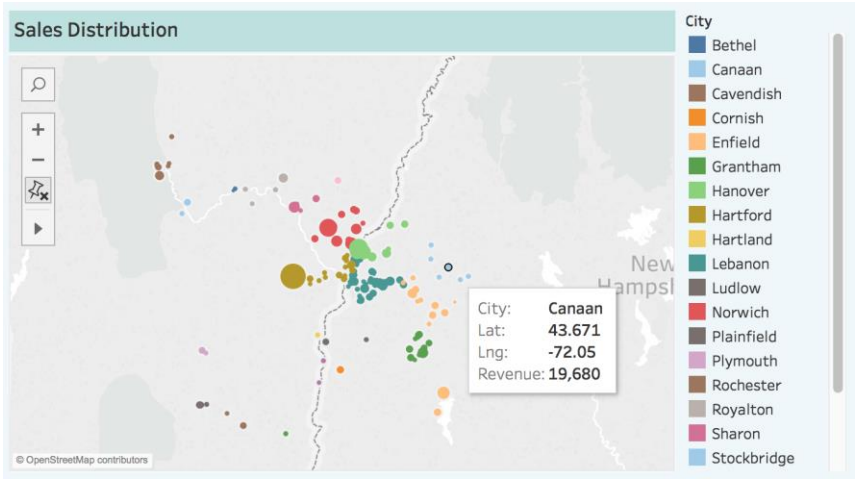


Figure 4.2.7

Overall, we design these two dashboards can give company the view of outside industry performance comparing to their own, and internal agent performance.

4.3 Predictive Analysis

We recommend a predictive model that can easily give a prediction to your target and provide a backup for future decision making. The general role of our model shown below:

- Exploit patterns found in historical and transactional data;
- Capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions;
- Guide decision making for candidate transactions;

To make the model useful, we decide to predict two metrics which are typically important for a real estate company.

Assessment Price should be provided to both company and customers, make it accurate is important for company to predict their revenue and satisfy the customers. **Days needed for selling the house** will let the company know the cost for maintaining it on the website and other potential cost.

As we explained in the former part, assessment price is so important in predicting revenue and remaining customers, and also predicting how long the house can be sold will be an additional information for future revenue prediction.

4.3.1 Predictive Analysis Model

Three different models are used to fit the prediction, and we try to find the model giving the best result.

- **Regression Tree**

This is a method that split the data into several parts, the target is trying to find the split part in which have the most similar data.

- **Random Forest**

A combination of several regression trees and each tree has some different random chosen features of the original ones. Using the average prediction as the result.

- **Artificial Neural Networks**

An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in a brain. Each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another.

4.3.2 Predictive Analysis Result

In our three models, only regression tree can give an interpretable result. Looking at the top part of the tree (Figure 4.3.1), we can find many information, Bedrooms – Total is the first feature to be split which means it is the most important to separate different properties by price, and similarly SqFtTotFn has the second importance. When understanding the entire tree, managers can know what is important for evaluating the price of a property.

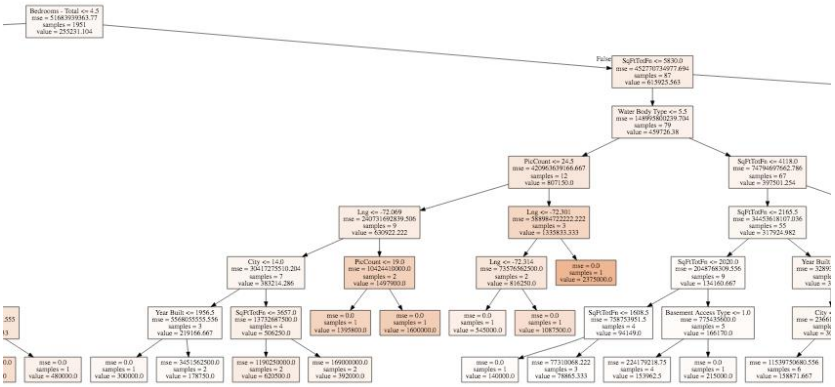


Figure 4.3.1

We define the accuracy of our price prediction as: The difference between our **assessment** and **the real sold price** is smaller than some percentage of the real sold price.

Method \ Tolerance	**10%**	**20%**	**30%**
Regression Tree	**24.73%**	**43.13%**	**58.06%**
Random Forest	**31.06%**	**55.68%**	**68.58%**
Artificial Neural Networks	**40.33%**	**63.27%**	**74.92%**

It is not hard to know all three models give a quite good result. Using these predictive model, we can predict the selling price with an 20% tolerant rate in 60%
Accuracy.

We define the accuracy of our days needed prediction as: The difference between our **assessment** and the **real days** is smaller than *n* weeks.

Method \ Tolerance	one week (14 days)	two weeks (28 days)	three weeks (42 days)
Regression Tree	**3.11%**	**5.97%**	**8.12%**
Random Forest	**3.35%**	**5.38%**	**7.53%**
Artificial Neural Networks	**13.33%**	**22.56%**	**31.43%**

Unfortunately, the result of days needed assessment is really bad with highest 31% accuracy and large tolerant rate. This can be led by different reasons, such as the low complexity of model and there even not a relationship between the days needed and our house features. A relationship between time period and days needed can probably be a potential factor. Although the result is not ideal, we can still know something useful, and based on these, company can gather more useful data and improve the model to make prediction more accurate.

Section 5: Implementation

5.1 Kotter's 8-step Change Model

Kotter's 8-step change model is widely accepted across all industries as an effective model for implementing organizational change. In this report, each step of the Kotter model is discussed using Jarboe Group, the real estate company, as the contextual example. Kotter's change model includes eight steps: create a sense of urgency, form a powerful coalition, create a vision for Change, communicate the vision, remove obstacles, create short-term wins, build on the change and embed the change in culture.

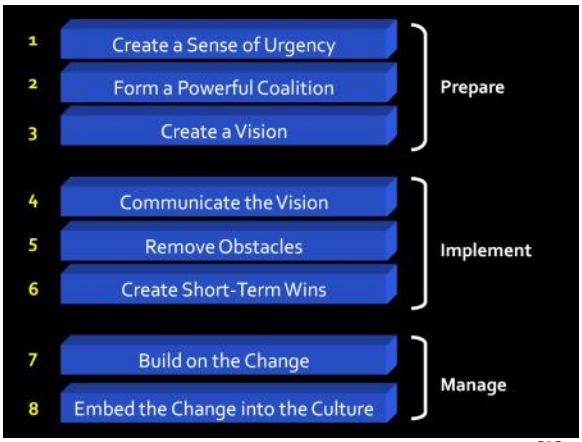


Figure 5. Kotter's 8-step Change Model [3]

- **Create a Sense of Urgency**

In the information era, adopting appropriate data warehousing (DW) could provide the company with immeasurable benefits, such as potential high returns on investment, competitive advantage, and increased productivity of corporate decision-makers. Unfortunately, our client, the Jarboe Group, only uses MS Excel for data analysis and does not use a data warehouse or business intelligence tools. Thus, we can see data warehouse as a way to improve Jarboe Group's technical infrastructure.

- **Form a Powerful Coalition**

Within Jarboe Group, the CEO, data analyst, and real estate agents are strong advocates for data warehouse and business intelligence tool.

- **Create a Vision for Change**

The CEO will generate the new vision: he/she would first create plan of business needs and then would evaluate expected benefits from the data warehouse.

- **Communicate the Vision**

The CEO will introduce the data warehouse and business intelligence tool to the employees in the strategic planning meeting. What's more important, the CEO will elaborate the importance and benefits of building a data warehouse and applying BI system.

- **Remove Obstacles**

The CEO should convince the employees that the benefits brought by DW are not only for the company but themselves.

- **Plan for and create short-term wins**

The benefits are expected greater than shown in the budget. The implementation of DW and BI tool enable the employees to improve efficiency and help the CEO to make better decisions.

- **Build on the change**

The Jarboe Group can use DW for financial analysis, marketing analysis and agent performance analysis. But the employees may not realize or agree how the DW can benefit the company.

- **Embed the changes into the culture**

The Jarboe Group should continue developing DW and figure out the appropriate BI tool for its business.

From the Kotter's 8-step change model discussed above, the challenges of BI implementation would be the following: firstly, the Jarboe Group has not used DW before. It would take a long time to build an unfamiliar DW. Secondly, the employees may be reluctant to give up the current working system. If employees are used to working with MS EXCEL, they may be unwilling to learn how to use BI tool. Based on the challenges may occur, we come up with several recommendations for the Jarboe Group. The first thing the Jarboe Group should do is to learn from other real estate companies to find out which kind of DW and BI tool would fit its needs. It would be easier to implement the BI if the CEO enable the employees to realize that the DW and BI tool could improve their performances.

5.2 Issues Related to Data Quality

Problems with data quality are costly to an enterprise. Understanding the root causes of common data quality issues is an essential first step for Jarboe Group to manage the information. The potential data quality issues are listed as below:

- **Incomplete Data**

As the data has not been entered in the system correctly or certain files may have been corrupted, the remaining data has several missing variables. Taking our dataset for example, some addresses do not include zip codes, which lead the remaining information can be of little value. The way to mitigate this issue is to check the completeness of data after entry to see if entered data meet current and future business information needs.

- **Duplicate Data**

Someone may simply enter the data multiple times by accident. Duplicate identification tools can help match and link records based on certain standards, allowing for greater uniqueness.

- **Inconsistent Formats**

Since data come from many sources, the data may be stored in inconsistent formats. Setting uniform concepts for data quality would be helpful to keep the data consistent.

- **System Upgrade**

There are chances of information getting lost or corrupt when the database is updated. The Jarboe Group should have backups of the initial data.

Section 6: Conclusion and Recommendations

6.1 Summary and Conclusion

We built two dashboards for our client The Jarboe Group, who is in the real estate business and help customers to buy and sell properties. The first dashboard Market Analysis is a strategic dashboard and can be used to: check key business data such as revenue, market share, number of closed sales and value of the max sales; compare sales performance and length of sales cycle with competitors; and also compare revenue and accuracy for estimating property value with the industry average. The second dashboard Agent Performance is a tactical dashboard and can be used to: check the revenue, number of sales and length of sales cycle of individual agents; compare sales performance and average sales cycle among the agents; check sales regions by agent; and also monitor the price per square foot in different regions. We also developed two predictive analysis models, one can be used to predict how soon a certain property can be sold, and the other one can be used to predict the sold price of a certain property, in other words, the model can help the agents to provide a better price estimation to their customers.

6.2 Recommendations

We recommend The Jarboe Group to build a data warehouse or at least a database to store their operational and transactional data, instead of using Excel which has limited space, will run very slow when the amount of data gets large, and is easy to accidentally delete or change data. To be able to stay ahead of their competitors and provide better managements to their agents, The Jarboe Group should adapt a BI system like the one we built, purchase a data analysis and visualization tool like Tableau, get their data analyst trained with the BI system so he can maintain and continue

to build/refine the system, lastly have their managers learn how to use the dashboards effectively. To make the predictive analysis models produce more accurate results, we recommend The Jarboe Group to collect more relevant data such as nearby school ratings, neighborhood safety, level of noise around the property, and so on. Finally, because The Jarboe Group is a small local real estate brokerage, we recommend them to subscribe Redfin datacenter to stay up to date of what data the real estate industry is collecting and how the data are used to grow business.

Reference:

[1] Performance Dashboards: Measuring, Monitoring, and Managing Your Business
By Wayne W. Eckerson: Charper 4 EXHIBIT 4.1 BI Maturity Model
[2] https://www.kaggle.com/samdeeplearning/real-estate-vermont-newhampshire/data
[3] https://thevirtualleader.files.wordpress.com/2011/12/change_8-steps.jpg?w=450&h=337

Appendix:

Appendix A1: Codes of Artificial Neural Networks

```
import numpy as np
import random
import tensorflow as tf
from sklearn import preprocessing
import pandas as pd

df = pd. read_excel (r'analysis.xls', sheetname=0)

le = preprocessing. LabelEncoder ()
for i in df. columns. values [1:8]:
    le.fit(df[i])
    df[i] = le. transform(df[i])
    df[i] = df[i]. astype('int')

#trainX = np. array (df.loc [0:1999,'Water Body Type':'DOM'])
#trainy = np. array(df.loc[0:1999,'Price'])
trainX = np. array (df.loc [0:1999,'Price':'Garage Capacity'])
trainy = np. array(df.loc[0:1999,'DOM'])
trainy = trainy. reshape (2000,1)

#testX = np. array (df.loc [1999:'Water Body Type':'DOM'])
#testy = np. array(df.loc[1999:'Price'])
testX = np. array (df.loc [1999:'Price':'Garage Capacity'])
testy = np. array(df.loc[1999:'DOM'])
testy = testy. reshape (np. shape(testy) [0],1)


x = tf.placeholder(tf.float32,[None,19])
```

```
y = tf.placeholder(tf.float32,[None,1])

W1 = tf.Variable(tf.truncated_normal([19,200],stddev=0.1))
b1 = tf.Variable(tf.zeros([200])+0.1)
z1 = tf.matmul(x,W1)+b1
h1 = tf.nn.relu(z1)

W2 = tf.Variable(tf.truncated_normal([200,50],stddev=0.1))
b2 = tf.Variable(tf.zeros([50])+0.1)
z2 = tf.matmul(z1,W2)+b2
h2 = tf.nn. tanh(z2)

W3 = tf. Variable (tf. truncated_normal ([50,1], stddev=0.1))
b3 = tf. Variable (tf. zeros ([1]) +0.1)
z3 = tf.matmul(h2,W3)+b3
h3 = tf.nn. tanh(z3)

#W4 = tf. Variable (tf. truncated_normal ([100,1], stddev=0.1))
#b4 = tf. Variable (tf. zeros ([1]) +0.1)
#z4 = tf.matmul(h3,W4)+b4

y_prediction = z3
#loss = tf.reduce_mean(-tf.reduce_sum(y * tf.log(y_prediction), reduction_indices=[1]))
#loss = tf.reduce_mean(tf.nn.cross_entropy_with_logits(labels = y, logits = y_prediction))
loss = tf. reduce_mean((y-y_prediction) **2)

Train = tf. train. GradientDescentOptimizer (0.000001). minimize(loss) #DOM:0.00001
#Train = tf. train. MomentumOptimizer (learning_rate=0.5, momentum = 0.9). minimize(loss)

init = tf. global_variables_initializer ()


#correct_p = tf. less_equal(abs(y-y_prediction), y*0.3)
correct_p = tf. less_equal(abs(y-y_prediction), 7)
accuracy = tf. reduce_mean (tf. cast (correct_p, tf. float32))

# using validation set to determine hyperparameters.
with tf. Session () as sess:
    sess.run(init)
    for epoch in range (1500): #DOM:650 #price:850
        sess.run (Train, feed_dict= {x: trainX, y: trainy})
        acc = sess.run (accuracy, feed_dict= {x: testX, y: testy})
        cost = sess.run (loss, feed_dict= {x: trainX, y: trainy})
        #print ('batch: '+ str(batch)+', node: '+str(node)+', learning rate: '+str(lr))
        print('Iter'+str(epoch+1) +', Training loss: '+str(cost)+', Accuracy: '+str(acc))
```

Appendix A2: Codes of Regression Tree

```
import pandas as pd
from sklearn. model_selection import train_test_split

df = pd. read_excel (r'analysis.xls', sheetname=0)

from sklearn import preprocessing
import numpy as np

le = preprocessing. LabelEncoder ()
for i in df. columns. values [1:8]:
    le.fit(df[i])
    df[i] = le. transform(df[i])
    df[i] = df[i]. astype('int')

data = df [df. columns [1:20]]
y = df['Price']
#y = df['DOM']
#y = y/(24*60*60*(10**9))

X_train, X_holdout, y_train, y_holdout = train_test_split(data.values, y, test_size=0.3,
random_state=17)

from sklearn. tree import DecisionTreeRegressor, DecisionTreeClassifier

reg_tree = DecisionTreeRegressor (max_depth=3, random_state=17)
reg_tree.fit (X_train, y_train)
reg_tree_pred = reg_tree. predict(X_holdout)
reg_tree. score (X_holdout, y_holdout)

from sklearn. model_selection import GridSearchCV
tree_params = {'max_depth': range (2,100), 'max_features': range (2,18,2)}

locally_best_tree = GridSearchCV (reg_tree, tree_params, cv=5, n_jobs=-1, verbose=True)
locally_best_tree.fit (X_train, y_train)
print ("Best params:", locally_best_tree. best_params_)
print ("Best cross validaton score", locally_best_tree. best_score_)
reg_tree = DecisionTreeRegressor (max_depth=7, max_features = 16, random_state=17)
reg_tree.fit (X_train, y_train)
reg_tree_pred = reg_tree. predict(X_holdout)
reg_tree. score (X_holdout, y_holdout)

i = reg_tree_pred - y_holdout
j = abs(i) <= y_holdout*0.3
#j = abs(i) <= 14
```

```
import numpy as np
np. shape(j)
sum(j)/np. shape(j)[0]

from sklearn. tree import export_Figureviz
import Figureviz

#dot_data = StringIO ()
dot_data = export_Figureviz (reg_tree, feature_names = df. columns [1:20], out_file=None,
filled=True)
Figure = Figureviz. Source(dot_data)
Figure

from sklearn. ensemble import RandomForestRegressor
rf = RandomForestRegressor (n_estimators=200, max_depth = 18, max_features = 6,
random_state=17) # you code here
rf.fit (X_train, y_train)

tree_params = {'max_depth': range (2,20), 'max_features': range (2,18,2)}

locally_best_tree = GridSearchCV (rf, tree_params, cv=5, n_jobs=-1, verbose=True)
locally_best_tree.fit (X_train, y_train)

print ("Best params:", locally_best_tree. best_params_)
print ("Best cross validaton score", locally_best_tree. best_score_)

RF_tree_pred = rf. predicts(X_holdout)
rf. score (X_holdout, y_holdout)

i = RF_tree_pred - y_holdout
j = abs(i) <= y_holdout*0.3
#j = abs(i) <= 21
import numpy as np
np. shape(j)
sum(j)/np. shape(j)[0]
list(RF_tree_pred)

from sklearn. tree import export_Figureviz
import Figureviz

#dot_data = StringIO ()
dot_data = export_Figureviz (reg_tree, feature_names = df. columns [0:22], out_file=None,
filled=True)
Figure = Figureviz. Source(dot_data)
Figure
```