NYC TAXI DATA-SCIENCE-ANALYTICS PORTAL

By: Jacky Ma (2015-09-06)

Charter Statement

The project is to build a **data-science-analytics portal for NYC Taxi** which aims to improve the utilization of taxi drivers and the satisfaction of passengers.

Project purpose

Why it is important?

NYC Taxi is facing fierce competition from Uber. Uber attracts a lot of passengers through its innovative service, but it will do no good to the passengers in the long run if Taxi has been totally driven out of the market leaving Uber as the monopolist. Therefore, it is important to help the taxi drivers to stay competitive to maintain a healthy market in order to maximize the benefits of customers.

What will it try to solve?

While competitiveness comprise of a lot of different factors, one of the major factors that the Taxi drivers are in significant disadvantaged is the difference in technology.

It is well-known that heavy use of data science analytics is critical for Uber to excel in operation efficiency and customer satisfaction. The project aims to narrow the "*technology gap*" by building a data science analytics portal for NYC Taxi, which will provide useful analytics for Taxi drivers to reduce their idle time and for passengers to better estimate their trip duration in advance.

What will be different when the project completes?

The portal, with its data science analytics engine, will provide prediction on:

- Demands for taxi at different locations
- Estimates of trip durations

The demand prediction can help taxi drivers to decide staying at or heading to locations where demand is high, in order to archive higher utilization. The estimates of trip durations can help drivers and passengers to have better time management, and results in higher passenger satisfaction. Both will contribute to the competitiveness of NYC Taxi drivers against the threat of Uber.

Project objectives and success criteria

Objective	Success Criteria
Build a website for NYC Taxi with cost no more than $12,000 and launch on or before 15 Nov, 2015	- A website shall be launch on or before 15 Nov, 2015. - Website expenses are $12,000 or less, including the cost of man hour of the team.
Generate useful data science analytics to NYC Taxi drivers and passengers	- Prediction of Taxi demand with respect to location, day of week, time and weather shall be provided. The prediction shall results in better estimation than the average Taxi demand during the hour (i.e. the base case). - Prediction of Trip duration with respect to pick up location and destination, day of week, time and weather shall be provided. The prediction shall results in better estimation than a distance-only prediction (i.e. the base case).
Attract users to create impact	- On or before Dec 31, 2015 the accumulate page view shall be over 5000. - Collect 50 or more feedbacks from Taxi drivers or passengers. - Over 5% of visitors indicate that they will re-visit the web for reference.

High-level requirements

- Taxi drivers must be able to see the predicted demand of Taxi in his proximity for this moment.
- Passenger must be able to get the predicted trip duration by inputting the pickup location and destination.
- The website must maintain a good layout on mobile browser (as well as on desktop browsers).
- Feedback channels must be provided.

Assumptions and constraints

- The website will be authored in-house to keep cost down.
- Raw data about NYC Taxi usage and NYC geo-data will be available.
- A usable prediction model can be trained within reasonable time.

High-level risks

- Raw data about NYC Taxi usage and NYC geo-data is unavailable or corrupted
- Unable to train a usable prediction model within reasonable time.
- Unable to attract users to visit the website.

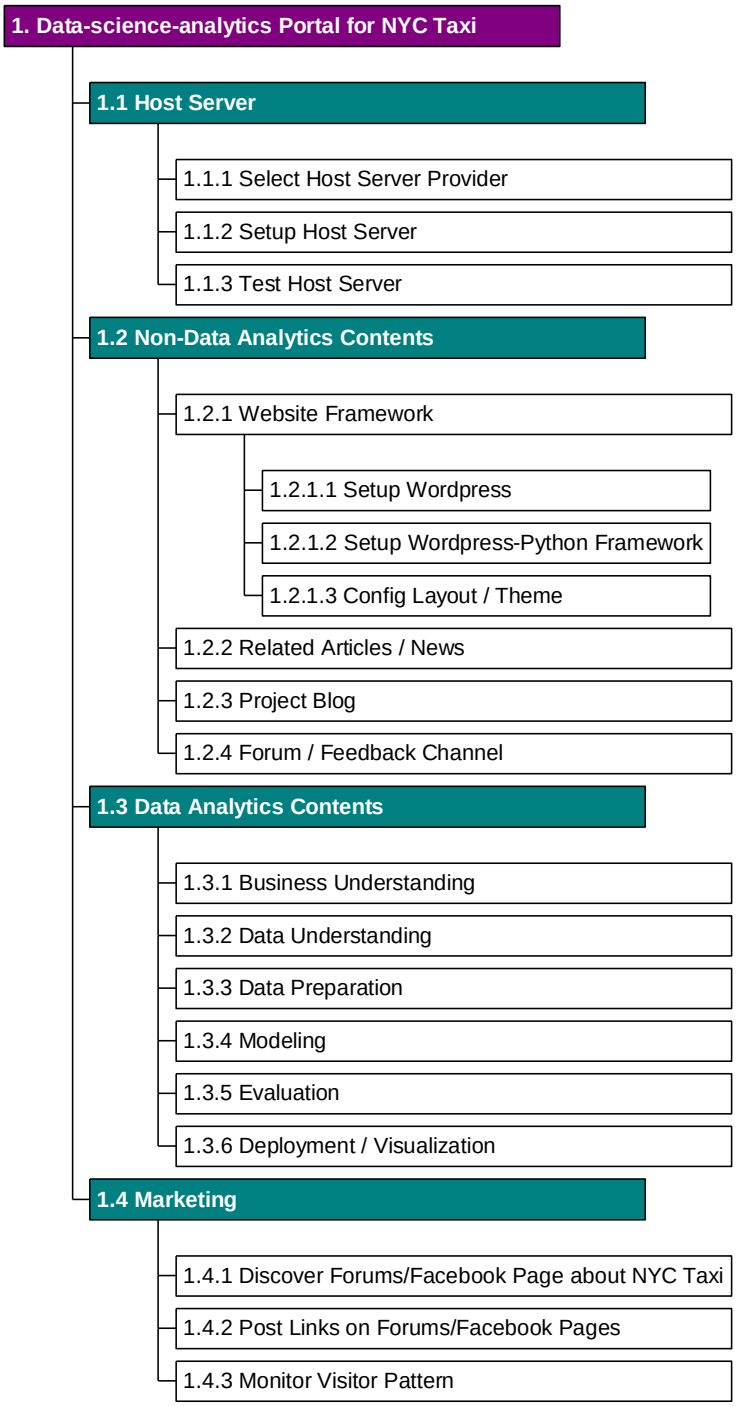
Summary milestone schedule

- Finish Testing the Host Server
- Finish Setting up the Server (WordPress + Python + Layout)
- Gather data needed for training the prediction model for taxi demand
- Gather data needed for training the prediction model for trip duration
- Finish the prediction model for taxi demand
- Finish the prediction model for trip duration
- Visualize the prediction model for taxi demand
- Visualize the prediction model for trip duration
- Launch website

Summary budget

Items	Cost
Labor - Web Editor	**$2520**
Labor - Data Scientist	**$8320**
I.T. Service or Software	**$350**
Total	**$11190**

Work Breakdown Structure



Sequence Project Activities

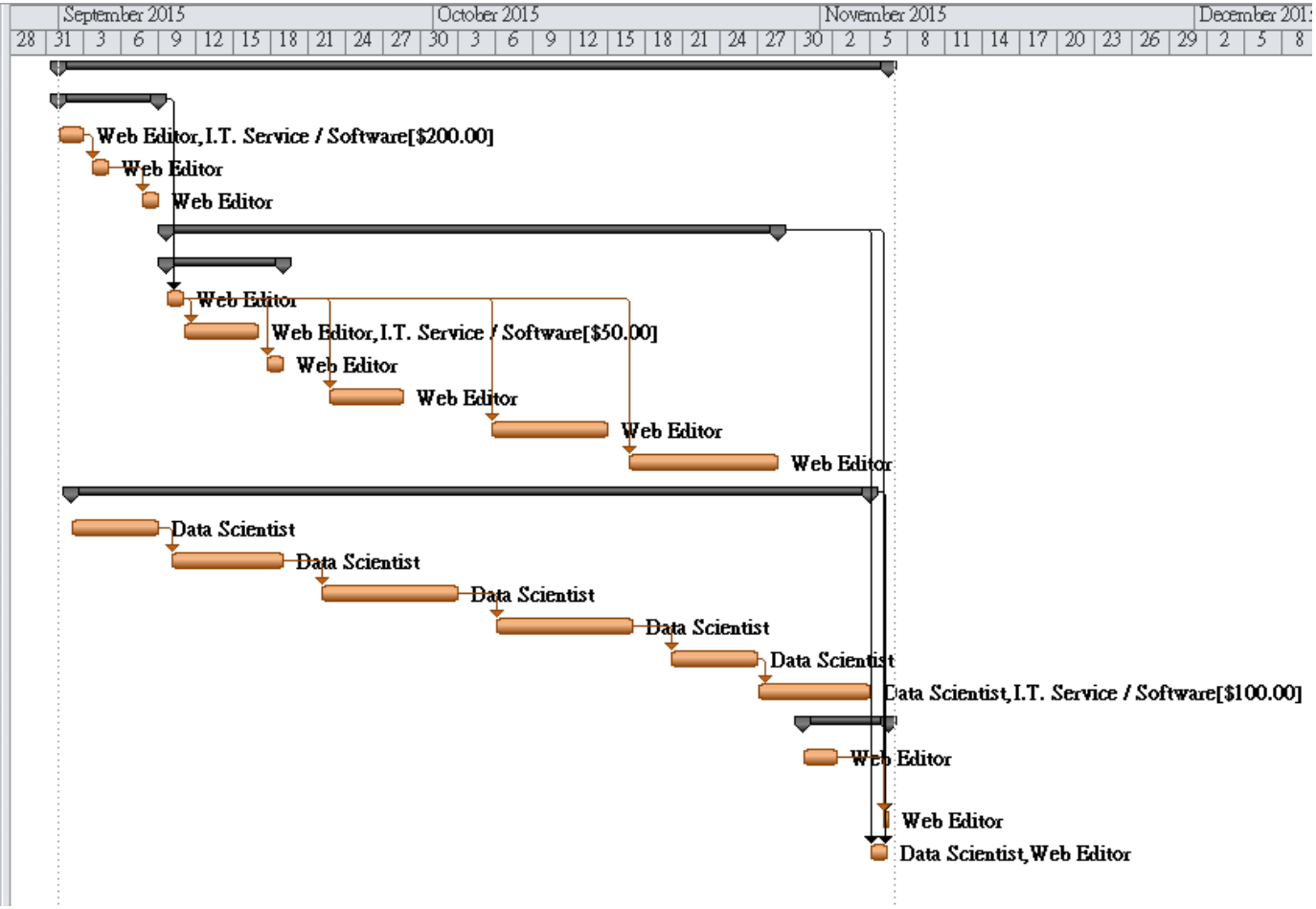
WBS ID	Activity / Task Name	Predecessor	Time Estimate
1.1.1	Select Host Service Provider	Nil	4 Hr
1.1.2	Setup Host Server	1.1.1	4 Hr
1.1.3	Test Host Server	1.1.2	4 Hr
1.2.1.1	Setup WordPress	1.1.3	4 Hr
1.2.1.2	Setup WordPress-Python Framework	1.2.1.1	8 Hr
1.2.1.3	Config Layout / Theme	1.2.1.1	4 Hr
1.2.2	Related Articles / News	1.2.1.1	8 Hr
1.2.3	Project Blog	1.2.1.1	16 Hr
1.2.4	Forum / Feedback Channel	1.2.1.1	16 Hr
1.3.1	Business Understanding	Nil	24 Hr
1.3.2	Data Understanding	1.3.1	32 Hr
1.3.3	Data Preparation	1.3.2	40 Hr
1.3.4	Modeling	1.3.3	40 Hr
1.3.5	Evaluation	1.3.4	24 Hr
1.3.6	Deployment / Visualization	1.3.5 & 1.2.1.2	40 Hr
1.4.1	Discover Forums / Facebook Page about NYC Taxi	Nil	4 Hr
1.4.2	Post Links on Forums / Facebook Pages	1.2.3, 1.3.6, 1.4.1	4 Hr
1.4.3	Monitor Visitor Pattern	1.2.4, 1.3.6	8 Hr

Project Schedule

As the team members are part-time based, Web Editor will commit 2 hrs on each work day (Mon to Fri) and Data Scientist will commit 4 hrs on each work day (Mon to Fri). The Schedule is arrange according to the dependency of tasks as well as the availability of resources (i.e. Web Editor and Data Scientist).

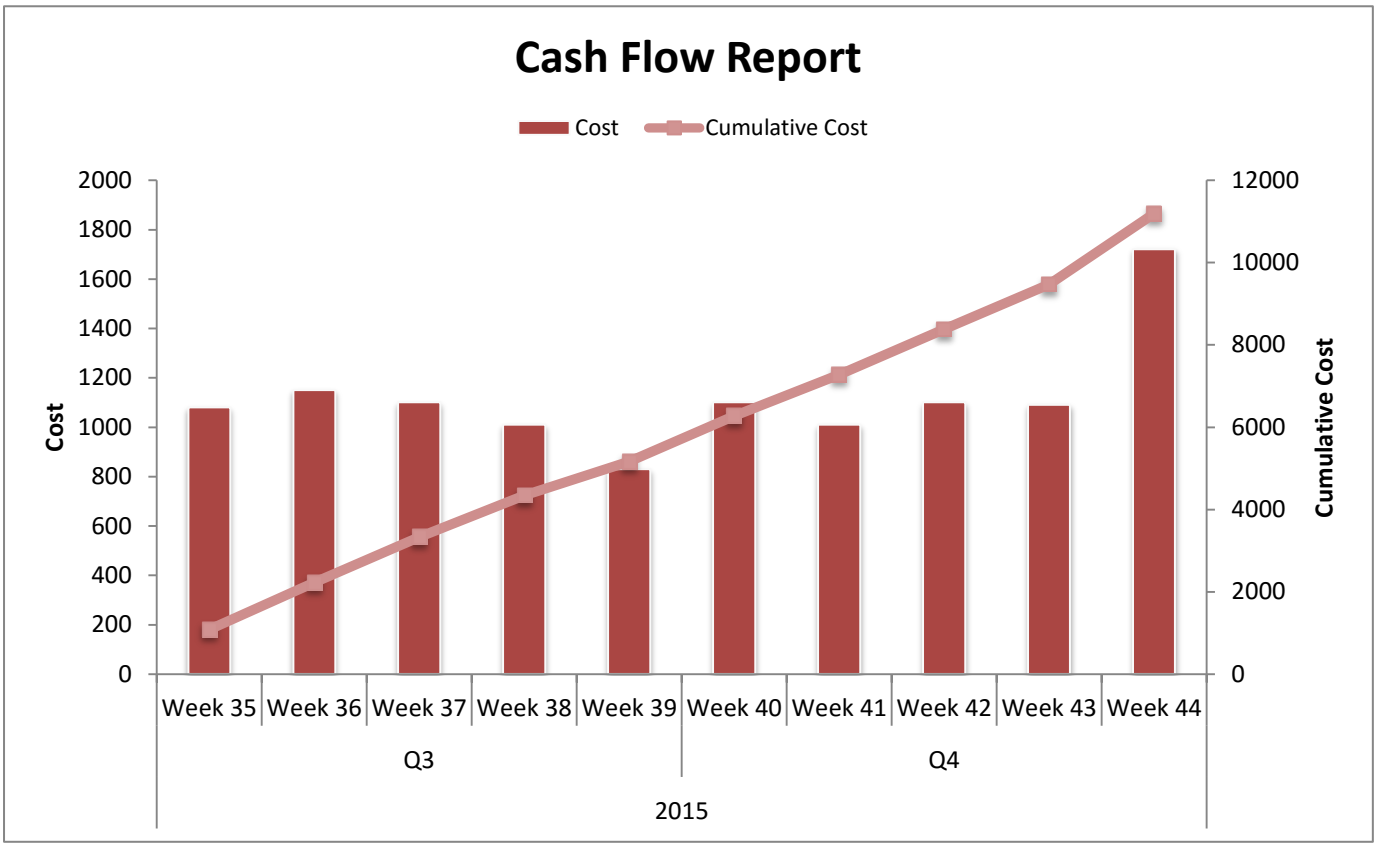
Task Name	Duration	Start	Finish	Predecessors	Resource Names
Data Science Analytics Portal for NYC Taxi	**292 hrs**	**Tue 1/9/15**	**Fri 6/11/15**		
Host Server	**12 hrs**	**Tue 1/9/15**	**Tue 8/9/15**		
Select Host Server Provider	4 hrs	Tue 1/9/15	Wed 2/9/15		Web Editor,I.T.
Setup Host Server	4 hrs	Thu 3/9/15	Fri 4/9/15	3	Web Editor
Test Host Server	4 hrs	Mon 7/9/15	Tue 8/9/15	4	Web Editor
Non-Data Analytics Contents	**56 hrs**	**Wed 9/9/15**	**Wed 28/10/15**		
Website Framework	**16 hrs**	**Wed 9/9/15**	**Fri 18/9/15**		
Setup Wordpress	4 hrs	Wed 9/9/15	Thu 10/9/15	2	Web Editor
Setup Wordpress-Python Framework	8 hrs	Fri 11/9/15	Wed 16/9/15	8	Web Editor,I.T.
Config Layout / Theme	4 hrs	Thu 17/9/15	Fri 18/9/15	8	Web Editor
Related Articles & News	8 hrs	Tue 22/9/15	Mon 28/9/15	8	Web Editor
Project Blog	16 hrs	Mon 5/10/15	Wed 14/10/15	8	Web Editor
Forum / Feedback Channel	16 hrs	Fri 16/10/15	Wed 28/10/15	8	Web Editor
Data Analytics Contents	**200 hrs**	**Tue 1/9/15**	**Wed 4/11/15**		
Business Understanding	24 hrs	Tue 1/9/15	Tue 8/9/15		Data Scientist
Data Understanding	32 hrs	Wed 9/9/15	Fri 18/9/15	15	Data Scientist
Data Preparation	40 hrs	Mon 21/9/15	Fri 2/10/15	16	Data Scientist
Modeling	40 hrs	Mon 5/10/15	Fri 16/10/15	17	Data Scientist
Evaluation	24 hrs	Mon 19/10/15	Mon 26/10/15	18	Data Scientist
Deployment / Visualization	40 hrs	Tue 27/10/15	Wed 4/11/15	19	Data Scientist,I.T.
Marketing	**24 hrs**	**Fri 30/10/15**	**Fri 6/11/15**		
Discover Social Media Related to NYC Taxi	4 hrs	Fri 30/10/15	Mon 2/11/15		Web Editor
Post Links on Social Media	4 hrs	Fri 6/11/15	Fri 6/11/15	6,14,22	Web Editor
Monitor Visitor Pattern	16 hrs	Thu 5/11/15	Fri 6/11/15	6,14	Data Scientist,Web Editor

	Task Name	Start	Finish
1	**Data Science Analytics F**	**Tue 1/9/15**	**Fri 6/11/15**
2	**Host Server**	**Tue 1/9/15**	**Tue 8/9/15**
3	Select Host Server I	Tue 1/9/15	Wed 2/9/15
4	Setup Host Server	Thu 3/9/15	Fri 4/9/15
5	Test Host Server	Mon 7/9/15	Tue 8/9/15
6	**Non-Data Analytics C**	**Wed 9/9/15**	**Wed 28/10/15**
7	**Website Framewor**	**Wed 9/9/15**	**Fri 18/9/15**
8	Setup Wordpress	Wed 9/9/15	Thu 10/9/15
9	Setup Wordpress	Fri 11/9/15	Wed 16/9/15
10	Config Layout / T	Thu 17/9/15	Fri 18/9/15
11	Related Articles & M	Tue 22/9/15	Mon 28/9/15
12	Project Blog	Mon 5/10/15	Wed 14/10/15
13	Forum / Feedback C	Fri 16/10/15	Wed 28/10/15
14	**Data Analytics Conte**	**Tue 1/9/15**	**Wed 4/11/15**
15	Business Understanc	Tue 1/9/15	Tue 8/9/15
16	Data Understanding	Wed 9/9/15	Fri 18/9/15
17	Data Preparation	Mon 21/9/15	Fri 2/10/15
18	Modeling	Mon 5/10/15	Fri 16/10/15
19	Evaluation	Mon 19/10/15	Mon 26/10/15
20	Deployment / Visua	Tue 27/10/15	Wed 4/11/15
21	**Marketing**	**Fri 30/10/15**	**Fri 6/11/15**
22	Discover Forums / Social Media	Fri 30/10/15	Mon 2/11/15
23	Post Links on Forum	Fri 6/11/15	Fri 6/11/15
24	Monitor Visitor Pattern	Thu 5/11/15	Fri 6/11/15



Project Budget

The main input of the project is labor. Labor cost account for 96.8% of total cost, whereas 22.5% for Web Editor and 74.4% for Data Scientist, under assumption that the hourly rate for Web Editor is $30 and for Data Scientist is $40. While both Web Editor and Data Scientist are our in-house member, the hourly rate is estimated by comparing to the market rate. I.T. Service and Software account for 3.1% of total cost, major expenses includes server hosting fee and software libraries for visualization.



Resource Name	30-Aug	6-Sep	13-Sep	20-Sep	27-Sep	4-Oct	11-Oct	18-Oct	25-Oct	1-Nov	Total
Web Editor	**$240**	**$300**	**$300**	**$210**	**$30**	**$300**	**$210**	**$300**	**$210**	**$420**	**$2520**
Select Host Server Provider	$120										*$120*
Setup Host Server	$120										*$120*
Test Host Server		$120									*$120*
Setup Wordpress		$120									*$120*
Setup Wordpress-Python Framework		$60	$180								*$240*
Config Layout / Theme			$120								*$120*
Related Articles & News				$210	$30						*$240*
Project Blog						$300	$180				*$480*
Forum / Feedback Channel							$30	$300	$150		*$480*
Discover Forums / Social Media Related to NYC Taxi									$60	$60	*$120*
Post Links on Forums / Social Media										$120	*$120*
Monitor Visitor Pattern										$240	*$240*
Data Scientist	**$640**	**$800**	**$800**	**$800**	**$800**	**$800**	**$800**	**$800**	**$800**	**$1280**	**$8320**
Business Understanding	$640	$320									*$960*
Data Understanding		$480	$800								*$1280*
Data Preparation				$800	$800						*$1600*
Modeling						$800	$800				*$1600*
Evaluation								$800	$160		*$960*
Deployment / Visualization									$640	$960	*$1600*
Monitor Visitor Pattern										$320	*$320*
I.T. Service / Software	**$200**	**$50**							**$80**	**$20**	**$350**
Select Host Server Provider	$200										*$200*
Setup Wordpress-Python Framework		$50									*$50*
Deployment / Visualization									$80	$20	*$100*
Total	**$1080**	**$1150**	**$1100**	**$1010**	**$830**	**$1100**	**$1010**	**$1100**	**$1090**	**$1720**	**$11190**

Responsibility Assignment Matrix

Activities / Roles	Project Manager	Data Scientist	Web Editor	Project Sponsor
Select Host Server Provider	A		R	
Setup Host Server	I		AR	
Test Host Server	I		AR	
Website Framework	I		AR	
Setup WordPress		I	AR	
Setup WordPress-Python Framework		I	AR	
Config Layout / Theme	A		R	I
Related Articles & News	A		R	
Project Blog	A		R	C
Forum / Feedback Channel	A	C	R	C
Business Understanding	A	R		C
Data Understanding		AR		
Data Preparation		AR		
Modeling		AR		
Evaluation	A	R		I
Deployment / Visualization	A	R	R	I
Discover Forums / Social Media Related to NYC Taxi	A		R	I
Post Links on Forums / Social Media	A		R	I
Monitor Visitor Pattern	A	R	R	I

Project Risks

Risk	Mitigation	Contingency	Impact	Likelihood	Overall Priority
If raw data about NYC Taxi usage and NYC geo-data is unavailable or corrupted, then the data science prediction model cannot be trainned	Mirrow the data source now	Use partial data that we already have	5	2	Medium
If a usable prediction model cannot be trainned within reasonable time, then the prediction function will be absent	Use adaptive development, such that we may find less powerful prediction model within reasonable time. And progress towards more powerful version afterwards.	'Downgrade' to visualization of past data instead of making prediction	5	3	High
If there are few users visit the website, then the website cannot make impact to the society	Cite more articles in the blog so that the site will more likely to appear in search engine	Submit the project in forums related to data analytics as well; create data visualization which may be more popular in social media	3	4	Medium
If a suitable hosting server cannot be found within budget, then the website cannot be set up	Use a lightweight model such that the requirement for data storage and processing power will be lower	Host the website in our own PC	5	1	Medium
If the team members cannot work on schedule as commited, then the project will fall behind schedule.	Weekends are unallocated and the last activity scheduled is two weeks before the deadline	Let team members work on weekends. Recruit more team members.	3	3	Medium
If the site is being hacked, then the website/programs have to be upload again	Install security patches. Backup the website and program	Reset the server and upload the website and program	3	1	Low
If the host server is being blocked by some search engine, then some users might not be able to visit the web.	Investigate the 'brand name' of hosting company when choosing server hosting packages	Relocate to another hosting company	1	2	Very Low

Risk assessment matrix