Contents

Onboard Revenue Prediction Assessment

Summary

This project uses data collected about a passenger during the purchase of a cruise to predict the spending of each passenger in the various departments onboard a cruise ship. This data does not include any personal information such as age, income or any other such information but only includes the purchase information such as purchase date, class of berth, price of ticket, length of cruise, destination and cruise date.

The customer is interested in planning for a project to predict the amount spent on the various departments onboard the cruise such as the spa, casino, bar and communications.

A range of information, such as the name of all features, have been anonymized to protect the proprietary information of our customer. Other sections have been edited to prevent confidential information being revealed. This is especially the case in sections which discuss the business impact of information.

Objectives

Business Objectives

This project aims to project the amount of revenue that will be generated in onboard revenue for a specific cruise. This overall objective includes several sub-objectives:

1. Project the total amount generated from each department to help the finance department with financial projections
2. Project the revenue generated to isolate the impact of marketing efforts from various campaigns
3. Identify customer personas with a strong propensity to consume services. These can then be target with marketing messages

<Edited> This section was considerably longer. In the customer report we discuss the impact of the predictions being made.

Technical Objectives

1. Understand the structure and nature of the data provided by the customer
2. Identify the quality of the data and provide suggestions about data quality improvement
3. Generate project implementation plan

Data analysis

Data Quality

We need to examine the quality of the data flowing into our machine learning models. Without accurate and complete data flowing into the system any machine learning model, no matter how sophisticated, will be unable to make accurate predictions. The quality of data is determined by factors such as accuracy, completeness, reliability, relevance and how up to date it is.

The tables below show some of the statistics which represent data completeness in the two tables which will be used as the major data sources for our machine learning model. Feature 8 in Purchasing History table and Feature 5 in Booking History data have significant data missing.

<Edit> There was a considerable discussion about the sources that discuss this missing data, but this is obviously sensitive information. We also discuss the specific predictions which would be affected, how improvements to this data completeness could improve the predictions and ideas for improving the source systems for this data.

Data Quality – Missing data

Purchasing History

(2,813,343 rows collected since 2018-01-01)

Features	Data type	Number of missing values	Missing rate (%)
Feature 8	object	278,804	9.9100
Feature 4	float64	25,482	0.9057
Feature 7	object	518	0.0184
Feature 2	int64	0	0.000000
Feature 12	int64	0	0.000000
Feature 15	int64	0	0.000000
Feature 13	int64	0	0.000000
Feature 1	float64	0	0.000000
Feature 14	int64	0	0.000000
Feature 3	int64	0	0.000000
Feature 6	object	0	0.000000
Feature 16	object	0	0.000000
Feature 5	object	0	0.000000
Feature 9	int64	0	0.000000
Feature 10	object	0	0.000000
Feature 11	object	0	0.000000

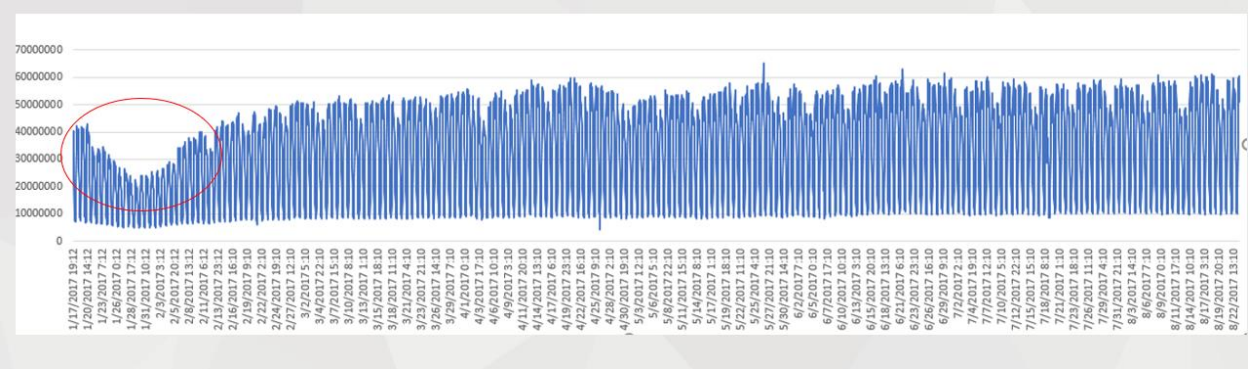
Booking History

(5,342,572 rows collected since 2018-01-01)

Features	Data type	Number of missing values	Missing rate (%)
Feature 5	object	5,123,055	95.8912
Feature 3	float64	85,902	1.6079
Feature 4	object	85,902	1.6079
Feature 9	float64	85,902	1.6079
Feature 2	float64	85,902	1.6079
Feature 8	int64	0	0.000000
Feature 1	int64	0	0.000000
Feature 7	object	0	0.000000
Feature 6	int64	0	0.000000

The chart below shows data volume change over time. Here we see that the data volume for a certain time range is much lower than other periods. A deeper investigation needs to be made into this issue.

- If this was a seasonality issue additional data could help to predict this e.g. holidays, weather
- If this was caused by data collection issue, we should avoid this issue happen in the future
- If this was caused by infrastructure changes, we need to understand these changes

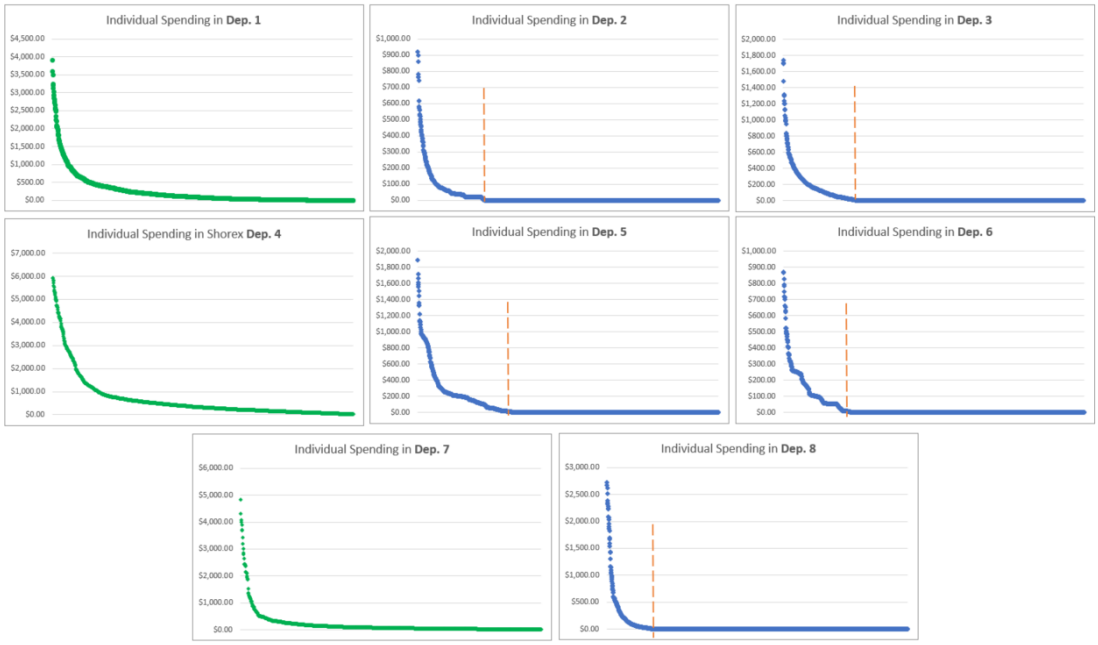


Data Patterns

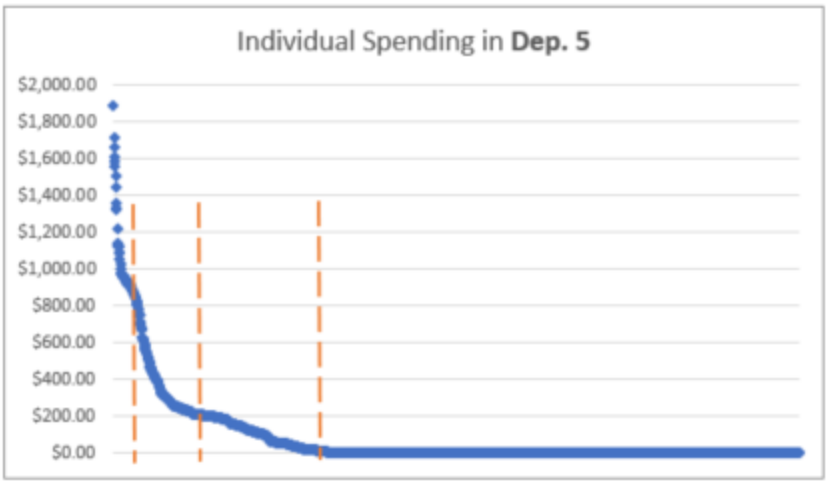
Data patterns are used as the basis of building a machine learning model. This also helps us to decide if any feature engineering is required, helps use to choose which algorithms to use and if we would get value from a multi-phased modeling approach.

The chart below shows how much each passenger spends in each department during a specific voyage. What we can see are:

- Quite a few passengers don't consume in some departments (Dep 2,3,5,6,8)
- All passengers spend money in some other departments (Dep 1, 4, 7)
- Obviously, we should use different approaches to do prediction for these two types of departments
    - I.e, identify who will spend in a certain department before doing revenue prediction.



We can see in department 5 there is a strong step function showing that it may be highly beneficial to try to group passengers using a clustering algorithm before we attempt to use a regression algorithm to predict their spend.

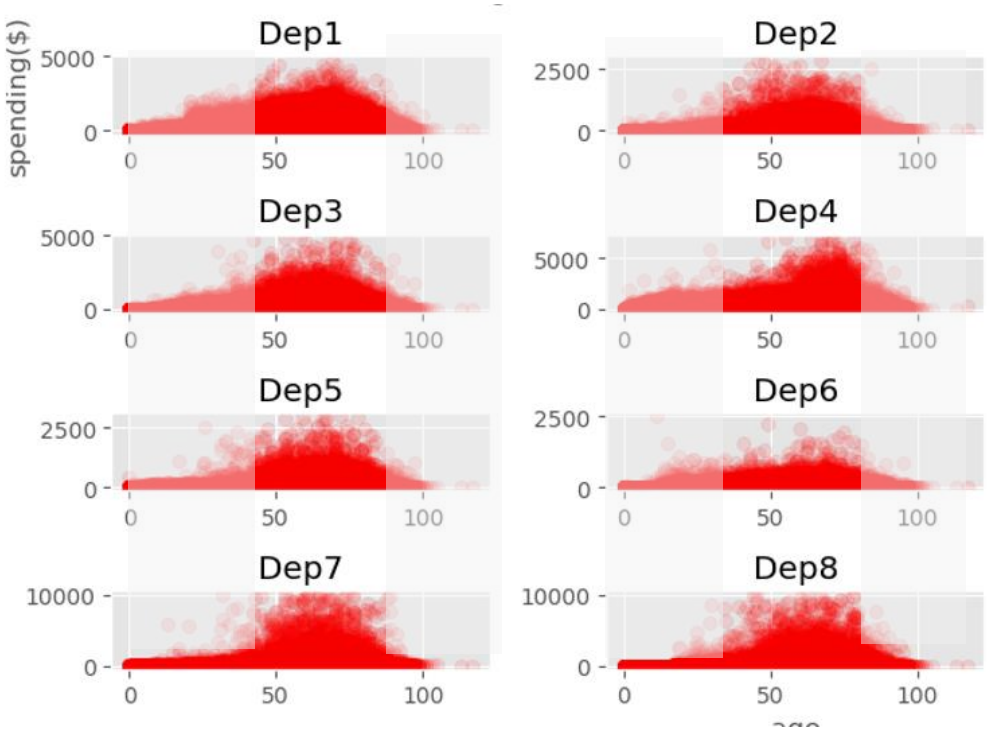


Data Relationship

We also investigated whether some features have strong correlation with the spending. And found:

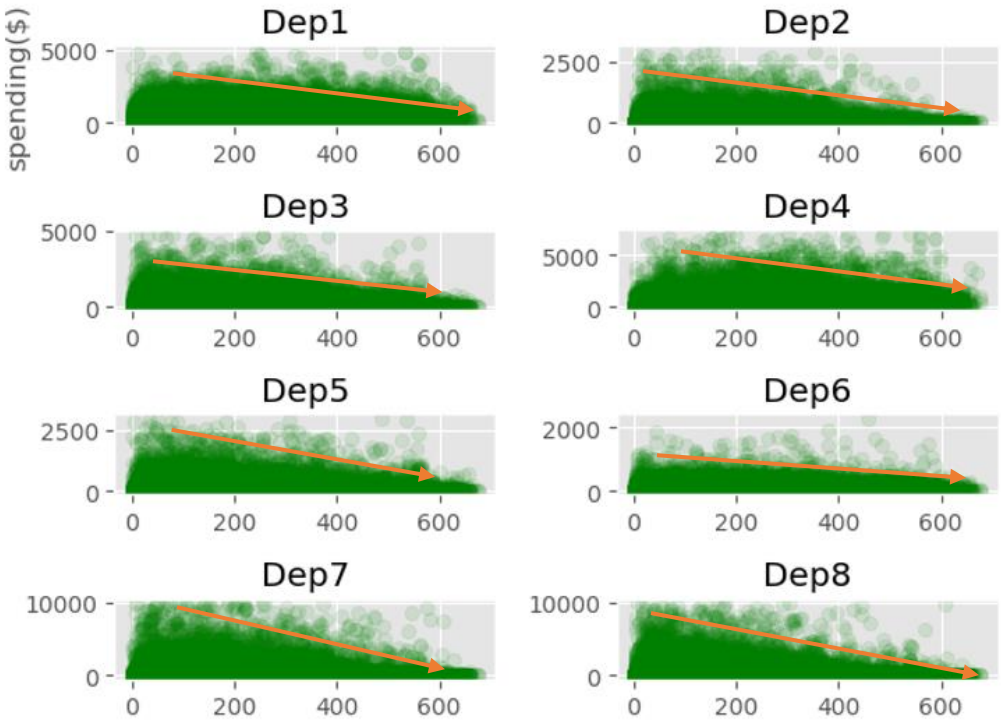
- Passengers whose feature 1 value between 40 and 80 spend more than passengers in other value ranges

Spending vs. Feature 1



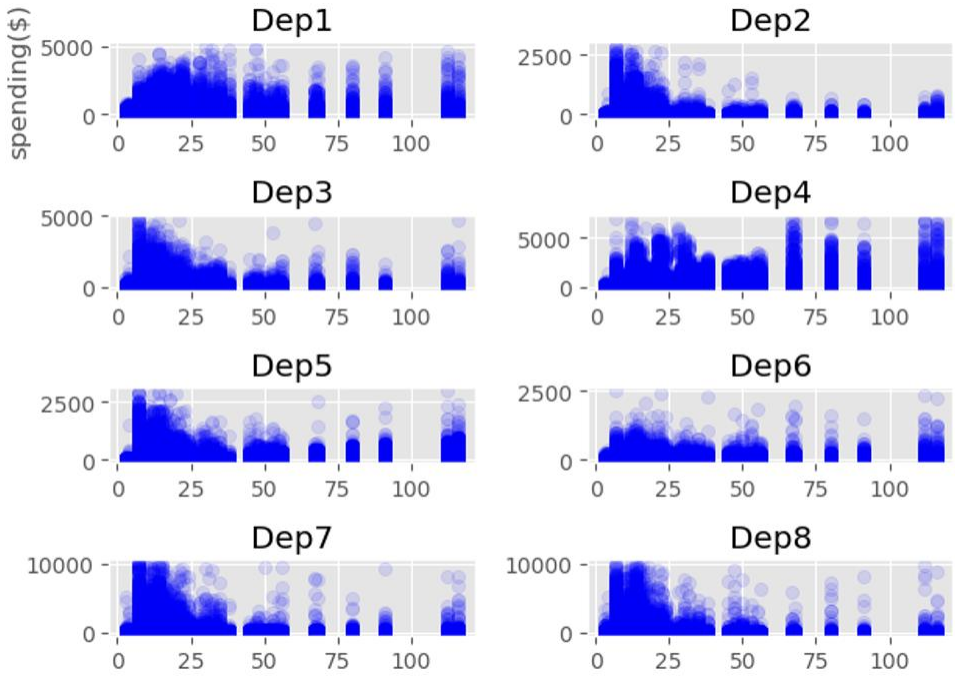
- Passengers with larger value of feature 2 might spend less.

Spending vs. Feature 2



- The relationship between passenger spending and feature 3 shows for centain values of feature 3, passengers might spend more

Spending vs. Feature 3



- Passengers whose feature 4 value equals 1 are more likely to consume in Dep 3, 4 and 7.
- Passengers whose feature 4 value equals 2 are more likely to consume in Dep 1, 5 and 6.

Spending records vs. Feature 4

- From the table below, we can see the revenue of some departments have stronger correlation:
  - Dep 2 and 3
  - Dep 1 and 6
  - Dep 6 and 7
- The revenues of some departments are irrelevant
  - Dep 4 and 8

Correlation between departments:

	Dep2	Dep3	Dep4	Dep5	Dep6	Dep7	Dep8
Dep1	0.531691	0.689818	0.583874	0.368323	0.71901	0.637478	0.44671
Dep2		0.788993	0.338431	0.258696	0.41283	0.614466	0.465876
Dep3			0.374995	0.376769	0.593573	0.692863	0.61764
Dep4				0.5559	0.687396	0.581012	0.120166
Dep5					0.696472	0.467944	0.352115
Dep6						0.712631	0.484007
Dep7							0.455195

Data Improvement Suggestions

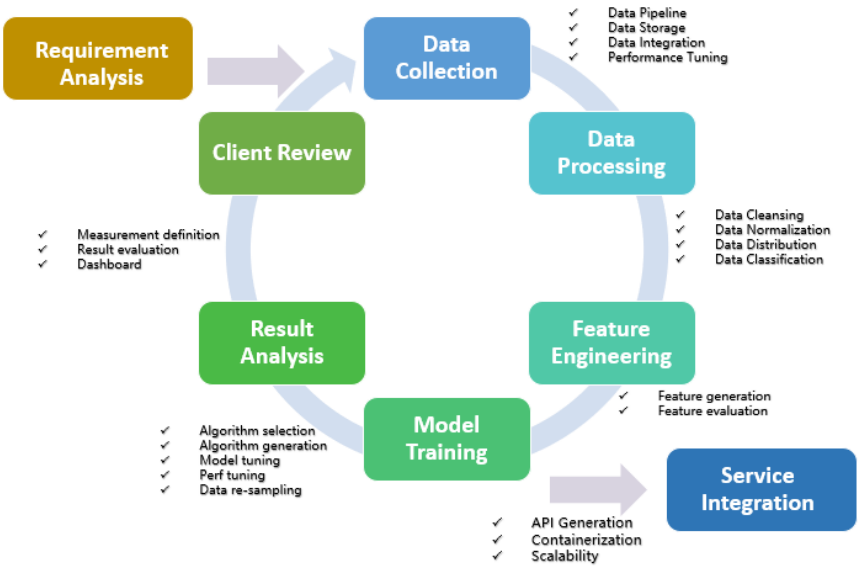
- Loyalty information is important for the revenue forecasting. Although we found patterns from the data provides. We still don't have the loyalty information in the data set. We would suggest the anonymized loyalty information is provided.
- Group booking information is also critical. The data we were provided did not provide the information about the travelling party a specific passenger is part of, the size of the party, number of children in that

party etc. This has a dramatic impact on spending in some departments. We suggest creating groups of types of traveling groups and then predicting spending after this.

<Edit> We had other suggestions for data improvements which are specific to our customer. We have left that information out of this report

Implementation Plan

The chart below shows the entire life cycle of the project implementation. Each section is discussed in detail in the previous and subsequent sections. Data Collection and Data Processing are discussed previously, Feature Engineering, Model Training and Result Analysis are discussed in the subsequent section.



Project Lifecycle

Feature Engineering

Based on our data analysis, we do see a lot of good features for building forecasting model. What we need to do are:

- Select right range of data with best data quality
  - Data of the last 12 months will be selected
  - Will keep data of U.S. market only because there is not enough data in other markets to train a good model
- Fill/remove missing data in original data set
  - All history data with “unknown” spending amount will be removed
  - All customers without associated customer Id will be removed
- Select the best features with the highest predictive power
  - PCA analysis will be used to select good features
- Generate additional features based on the original data set
  - Calculate the days between booking dates and departure dates

    - Calculate the number of passengers travel together
- Normalize feature values
    - Convert cabin type to rank number sorted by ticket price
    - Use mile as the unit of the distance if different units are applied
- Data sampling and generating training set and test set
    - 5% data will be randomly selected from the raw data to reduce the volume of data for analysis
    - 70% sampled data will be used as training set, and 30% as test set

Algorithm Selection

We need a two-step prediction solution：

1. Classification algorithm:

   Classify passenger for each department based on their consumption patterns. The reason is that if a significant number of passengers are not spending in a certain department, forecasting their spending is meaningless. We should use machine learning to predict the passengers who are likely to spend and doing spending forecasting against them in the next step.

   Logic regression algorithm with be used to do multi-class classification

2. Regression algorithm:

   Forecast spending of each passenger in each department. In this step, to get high performance and accuracy, only passengers who are considered to spend money in the certain department will be served into our machine learning model.

   Liner regression will be used to do the forecasting.

Model Tuning Plan

We will develop a model tuning plan in concert with the operational team. This is especially important when efforts are being put forth to use the predictions to affect to behavior of customers. This is also important when we are looking to make improvements to the data sources.

There is also a strong ability to use these predictions in different applications such as marketing where the goals of the model would be dramatically different than for finance purposes for example. Each of these models would need to be trained separately.

Projected Results Analysis

We have omitted this section from the report. This is highly sensitive information which our customer does not want revealed. This section also discusses the business impact of our low, medium and high accuracy predictions which is also very sensitive.

Service Integration

RESTful APIs will be provided to get input data and output machine learning results. Another set of RESTful APIs will be provided for audit (check how many data will processed, how many failures happened during prediction, etc.)

The service will NOT store data permanently. All the raw data for prediction and prediction results will be temporally stored by machine learning service for no more than 24 hours, such that client can retrieve the data in case they have problems to receive prediction results in time.

The server running machine learning service will be ran in clients' subscription on Azure.

Resource Plan

1 Data Engineer – Data Processing (30 person-days)

Skills: Hadoop, Data Lake, SQL Server, Streaming service

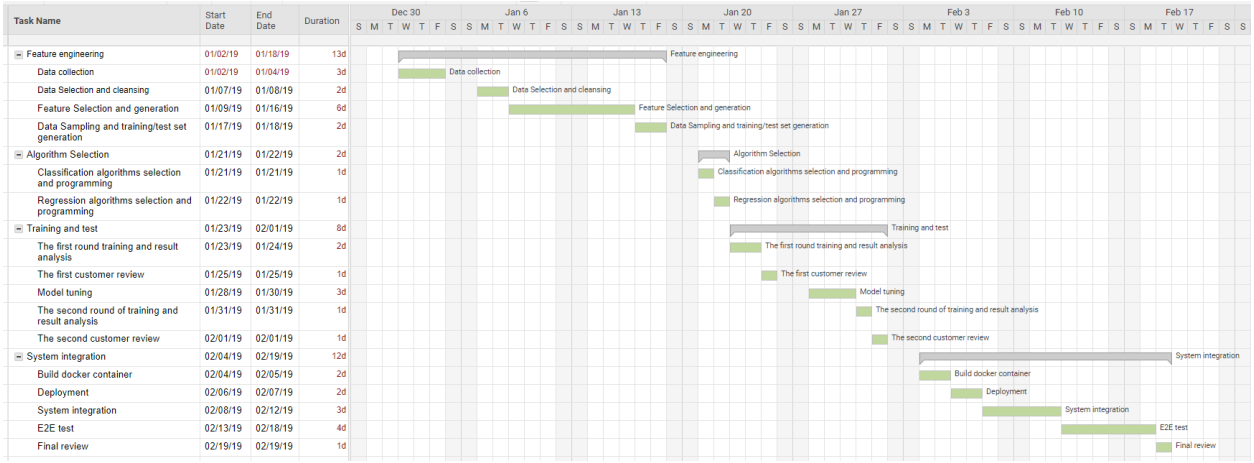
1 Data Scientist - Machine Learning (60 person-days)

Skills: Machine Learning, Deep Learning, Python, TensorFlow

1 Developer – System integration (40 person-days)

Skills: C#, ASP.NET Core, JavaScript, UI design, API design

Project Timeline



Task Name	Start Date	End Date	Duration
Feature engineering	01/02/19	01/18/19	13d
Data collection	01/02/19	01/04/19	3d
Data Selection and cleansing	01/07/19	01/08/19	2d
Feature Selection and generation	01/09/19	01/16/19	6d
Data Sampling and training/test set generation	01/17/19	01/18/19	2d
Algorithm Selection	01/21/19	01/22/19	2d
Classification algorithms selection and programming	01/21/19	01/21/19	1d
Regression algorithms selection and programming	01/22/19	01/22/19	1d
Training and test	01/23/19	02/01/19	8d
The first round training and result analysis	01/23/19	01/24/19	2d
The first customer review	01/25/19	01/25/19	1d
Model tuning	01/28/19	01/30/19	3d
The second round of training and result analysis	01/31/19	01/31/19	1d
The second customer review	02/01/19	02/01/19	1d
System integration	02/04/19	02/19/19	12d
Build docker container	02/04/19	02/05/19	2d
Deployment	02/06/19	02/07/19	2d
System integration	02/08/19	02/12/19	3d
E2E test	02/13/19	02/18/19	4d
Final review	02/19/19	02/19/19	1d

Budget

Small: ($15,000)

Goal: Training machine learning models with high performance and accuracy. Prediction can be done manually using trained model

Medium: ($50,000)

Goal: Training machine learning models with high performance and accuracy. Solution is wrapped up as a service with well-designed input/output APIs.

Large: ($100,000)

Goal: Training machine learning models with high performance and accuracy. Solution is wrapped up as a service with well-designed input/output and audit APIs. Seamlessly integrated with customer current data system. Full functional management UI.

Future Improvements

1. Retrain and maintenance:

   Machine learning Models should be retrained automatically every 3 months. And the trained model should be online to replace the old model after A/B test

2. Scheduled execution:

   The prediction will be kicked off manually in this project. But in the long run, the prediction can be scheduled, and executed automatically.

3. Realtime prediction

   This project will achieve batch prediction only. Realtime prediction will be enabled in future release,