Machine Learning Engineer Nanodegree
Capstone Proposal
Vaishali Sutaria
March 8, 2018

Proposal
This proposal is on cycle time in days, it takes for a new construction sites in a telecom industry given the jurisdiction implications.

Every project that is released for construction has a delivery date of completion. The company generally spends hundreds of dollars investing and forecasting on a new site build project based on the delivery date however, some of the major milestones take lot longer in cycle times and sometimes unpredictable which delays the build.

This is troublesome because it is like an open check book and open risk. In order to reduce the uncertainties of the delivery date and spending additional funding and resources I have this proposal. I would like to propose that the telecomm companies invest ahead of time for difficult jurisdictions or increase the delivery date. In order for such recommendation to be made we need to quantify and look at historical data on the performance in the different jurisdictions and classify them. This will allow the company to direct the resources accordingly. This will also help improve the forecast accuracy for finance team on the New Site Build Projects.

In this project I would like to analyze a fictitious dataset where I will analyze the days taken to complete 10 major milestones(features).

The goal is to finish the build in 2 years, I am using 730 days for this project. However not everything gets build in 2 years so logically we should provide funding ahead of time for those jurisdictions.

I want to use what I have learned in this program and identify the important features based on cycle times.

1. Create customer segments based on the available data
    a. I want to identify which cities / jurisdictions we should either increase/adjust the delivery date or we should provide funding in advance.

Here are the 10 major milestones (features) I would like to review:

Days to Vendor Handoff	MS1
Zoning Submitted	MS2
Zoning Approved	MS3
Candidate Identified	MS4
Candidate Approved	MS5
Site Acq /Jurisdiction Approved	MS6
CX Start	MS7
Tower Ready	MS8
CX Complete	MS9
ON AIR	MS10

Steps to achieve this:

1. Load Data
2. Explore Data
3. Prepare Data
   a. Normalizing the data
4. Naïve Predictor Performance
5. Import 3 supervised Learning Models to compare
6. Choose the best model and identify feature importance’s
7. Implement Model Tuning
8. Final Model Evaluation

Domain Background

This project is primarily keeping in mind the telecomm industry. The time it takes to build new sites in any given city. For the purposes of this project I will use 2 years (730 days) as a benchmark to complete the new site build after funding. However, building new sites does not always happen within 2 years.

Not building in expected time frame has various implications. The cost to build the sites presumably is higher if it takes longer because the cost of materials is expected to rise. Secondly, the coverage or capacity need of the company is not met in time which hinders the company performance. Ultimately, the wireless customers do not get to enjoy the benefits that the company want to provide because of the build delays.

I am interested in this project because once I have this model of machine learning build I can then apply this model to real world data from my company and use it to propose a change in the funding strategy.

Problem Statement

The problem is building a new site is taking lot longer than expected. Ideally, 2 years is the benchmark to build new sites for a telecomm industry. However, the reality is not all projects released to build in 2 years are build in time. There are sites that will build before time and there are sites that will build after the delivery date. The delay and acceleration of the project delivery has a financial impact. The fundamental way of funding strategy needs to be changed by jurisdiction not by cycle times. I want to identify by segmenting the customers which cluster should be given funding to begin the pre funding jurisdiction process. This approach will help ensure the New site is delivered on or before time.

Datasets and Inputs

I have used a randomly generated dataset of 579 samples with 10 milestones for every key. Using my core knowledge in this area I have created all possible situations that can occur. Sometimes the situations or milestones are skewed because of database issues and sometimes the milestones are skewed because of the delay.

Key – Key is a record created to represent a New Site Build
City – City
County – County
State – State of build
Zip – Zip code

Days to Vendor Handoff	MS1
Zoning Submitted	MS2
Zoning Approved	MS3
Candidate Identified	MS4
Candidate Approved	MS5
Site Acq /Jurisdiction Approved	MS6
CX Start	MS7
Tower Ready	MS8
CX Complete	MS9
ON AIR	MS10

Solution Statement

From industry standards I would say the benchmark model is if the site builds in 730 days no prefunding required. If the site builds take more then 730 days, then pre funding is required.

I am not quite sure which base model I will use for benchmark yet. Some of the ones I am considering are:
R2 Score
Naïve Predictor
Decision Tree regressor

Evaluation Metrics

I am going to use the projects I did in the ML program as a guide to for the evaluation metrics.

First I will begin evaluating the performance by learning curves or some other comparison. Not sure on the final strategy yet. One idea is comparing the number of training sets and accuracy by using decision tree model with maximum depths to ensure I am not under fitting or over fitting. If high bias, it will pay very little attention to the data and simplify with high error rate, if high variance then the model wont be able to generalize well.

Project Design

Steps to achieve this:

1. Load Data
2. Explore Data
3. Prepare Data
    a. Normalizing the data
4. Naïve Predictor Performance
5. Import 3 supervised Learning Models to compare
6. Choose the best model and identify feature importance's
7. Implement Model Tuning
8. Final Model Evaluation

This project is similar to the customer segmenting project we did in the program. I am going to use that as a model to help me walk through the steps of loading the data, exploring the data. Looks for skewed performance and normalize it. Want to see the feature distribution and the data correlation between the different features (10 milestones). If I find that the data is not normally distributed then I will apply feature scaling, using natural logarithm.

Once at this step will try to find any outliers in the data. (I suspect this is possible in real world data can be skewed because of database loading errors, transfers etc.) I don't want the noise so will remove those outliers.

After the data has been scaled to a more normal distribution and outliers removed, I can then apply PCA to understand which dimensions about the data best maximize the variance of the features involved. Will report the explained variance by dimensions.

Next I would try to reduce the dimensionality if possible.

Lastly I want to create clusters. I will use Silhouette coefficient for data points measure. I learned in the class that calculating the mean silhouette coefficient provides a simple scoring method of a clustering. Then I can see the results of clustering and see how my samples line up. I know that the ON AIR milestone above 700 should be classified differently then the ON AIR milestones below 700. Given this knowledge I will be able to see if the algorithm helps predict the clusters accurately.