AN INTELLIGENT METHOD AND SYSTEM TO PREDICT THE SERVICE VALUE BASED ON GEO LOCATION

Arhath Kumar[1], U Sudarshan Karanth[2]
[1]Assistant Professor, [2]Student
Dept of MCA, NMAMIT, NITTE

Abstract

In this paper, we focus on the problem of service cost prediction for a given service based on the geolocation. We, 20th century people lives in the era of Big-Data. The increased availability of huge amount of historical data and need to perform precise and errorless prediction of future behavior in several areas of services. Each service is unique and the cost of the service is dependent on the type of the service and the location of the service. The difference in the service charges for any given service depends on various aspects like supply and demand, quality of service and location. The information on the service cost was extracted from different sources. We predicted service cost for a given service at a given location. We used linear regression model. We predicted the service cost with error rate 0.087. and we are going to present the results for the above algorithm. The service charge prediction algorithm utilizes the price of a given service at various locations in the vicinity of target location to predict the service charge. The algorithm would be designed to accept the type of service, location and its value. Based on the input data the algorithm uses the service value at various locations in the vicinity of the location where the service value has to be determined, to predict the dynamic value of the service based on statistical model.

I. INTRODUCTION

Although the technology has improved the way the services are being delivered, there is enormous number of services for which the service charge/fee is not determined based on geo location. Each service is unique and the cost of the service is dependent on the type of the service and the location of the service. For example, a service charge for any service in a tire-1 city would cost an individual differently from that of an individual living tire-2 city. Also, the service charges for any given service at two different location within the tire-1 city is not the same.

For example, let us consider the factors for variation in value of housekeeping service at various locations with in a city. The house keeping services has huge demand in and around industrial areas of the city and because of which the value of the service is inflated at these locations when compared to other locations within the same city.

Although the technology provides better connectivity and access to various services at the fingertips, which helps in finding the right service provider. The value of these services is not dynamically determined based on the location and demand. As we grow up, we expect 'everything nearby' and also our smart phone to serve the relevant local results.

The prime of internet helped us to closely link with people and content otherwise would have been inaccessible. As the technology improved we moved on to web where we make use of email communication and social media. The main moto of web 2.0 is to engage, collaborate and share the information with each other.

The barter system might have been the earliest form of sharing. From sharing content and media online, Internet has now enabled people to share physical things like homes, vehicles, appliances, furniture, etc. with each other. Many example come into picture here.

- Crowd funding
- Apartment / House Renting and Couch surfing.

- Ridesharing and car sharing

With the current trend, mobile-centric universal commerce, vendors are increasingly making the effort to drive their presence globally through multiple channels by delivering the services of greater quality, to get more control on their businesses. One key area which is steady and consistently high development is hyperlocal market. Hyperlocal refers to very specific area, area in proximity of our home or business or current location.

Why hyperlocal?

Hyper local is the next gen marketing, [1] which combines both online and offline platforms in order to gain massive scale demand and deliver the goods in the shortest possible time.

Some of the current hyper local technology service providers are UBER, OLA. But, all these service providers they have their own service cost for their service, and customer can't able to set the price for the service.

And in some of the other technology drivers e.g. WYKE which works in hyperlocal market, customer can able to negotiate the price for the service. Our paper is present the idea of building a product which will be useful for predicting the service value of the service at a geo location. The cost [2] of the service is dependent with quality, supply and demand. Moreover, it is strongly dependent with location of the services. In this paper, we predicted service cost for a given service at a given location. We used linear regression model and we going to present the results for the above algorithm. In addition, we will discuss more about our approach.

II. PROBLEM STATEMENT AND DATASET

The convectional or orthodox method of determining the value of a service is dependent on the past experience of the person who requires the service. But the cost on any given service is dependent on various factors like location, inflation, supply and demand, and the conventional methods doesn't consider these factors while determining price of any given service. Also, currently there is no system to predict the service value of given service in the market. The features used for the prediction are same across the entire dataset, the target to be predicted can have subtle differences depending on the intended usage of the prediction. Data is collected from different sources and also, we generated some of the statistical data.

The dataset has approximately 12500 samples and 7 various features. The features include latitude, longitude, pin code, city, area popularity score (lsv), supply to demand ratio (suptodem) and price. The geographical location is an account for the spatial and temporal trends in prices. We incorporated the rankings to the location based on the popularity of the location. Using these features, it reduced the prediction error greatly.

III. METHODS

1. Data pre-processing

The data parsed from raw csv file may have mistakes both in original record as well as initial preprocessing. Outliers are determined by the human inspection of values and subsequently removed/corrected. Few of the examples have missing values and these are excluded for further analysis. The services with popularity score less than 2.5 were not included in the analysis since they are likely to be errors.

2. Database and Features

For the prediction of price for the service we mainly dependent on the location of the service, the popularity of the area and supply and demand ratio in that location. For the time series forecasting of information we represented year as categorical variable.

Initially the database is built with 12500 samples. For building of sample, we took the GPS co-ordinates of various locations in Bangalore area, then we pass this location details to Google's pin code finder api, which will retrieve the pin- code and city for the particular GPS co- ordinates. We have manually assigned rating for each of the pin-code. Using this rating we will set the area popularity score (lsv) for each of the samples. After the popularity score is assigned based on the timestamp we will generate a random number with in the range of 0.01 to 1.5 this will be assigned to the supply to demand ratio (suptodem). We took a random number within range of minimum and maximum salary (price) of a particular job, to set the price.

The data for time series forecasting including the features like date, maximum price and minimum price to forecast the price of the service. And we

felt number of samples in our dataset was sufficient of the training of our regression model.

IV. IMPLEMENTATION

We selected Linear Regression for the price prediction. The algorithm is implemented using Python's scikit-learn library [6].

To build a better base line performance with linear classifier we opted Linear regression model. The main goal is to minimize the sum of squared errors. The linear regression model fits a linear function to a set of data points. The function in form of:

$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots + \beta n * Xn$$

Where Y is the target variable which is the price and X1, $X2$, ... Xn are the predictor variables and β1, $\beta 2$, … βn are the coefficients that multiply the predictor variables. $\beta 0$ is constant X0 = 1 . n is the number of features.

Formally, the model for multiple linear regression, given n observations, is [3]

$$yi = \beta 0 + \beta 1 xi1 \beta 2 xi2 + \beta .. pxip + \varepsilon i \text{ for } i = 1,2, \ldots n.$$

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates $b0$, $b1$, ... bp are usually computed by statistical software.
The values fit by the equation $b0 + b1xi1 + ... + bpxip$ are denoted $\hat{y}_i$, and the residuals e_i are equal to $y_i - \hat{y}_i$, the difference between the observed and fitted values. The sum of the residuals is equal to zero.

$$\frac{\sum e_i^2}{n - p - 1}$$

The variance σ^2 may be estimated by $s^2 =$, also known as the mean-squared error (or MSE).
The estimate of the standard error s is the square root of the MSE.

For time series forecasting we used Autoregressive Integrated Moving Average Model (ARIMA Model) [4]. ARIMA is a statistical model for analyzing and forecasting time series data.

Its acronym itself is descriptive. **AR**- it uses dependent relationship between an observation and some number of lagged observation. **I**- it uses of dissimilarities of raw observation.**MA**- it uses dependency between an observation and residual error moving average model applied to lagged observation.

The notation for ARIMA is ARIMA (p, d, q)

Where p - lag order (count of the observation). d - count of raw observation differed. q - is the size of the moving average window. This time series forecasting is achieved using Python's Pandas [7] and Statemodels library.

To showcase the output, we built web interface. This web interface is designed to input the location details and type of service. After the input the request has been sent to the server. In the server, we kept running our algorithm which is trained with our training dataset. Once the request is received we run through this algorithm to predict the price for the given service at the specific location. A heat map is generated with price of the services at various location, and its, with different colors to differentiate among the various price range.

V. RESULTS

The results are shown below. For the given input at the location with the GPS coordinates 12.97551995, 77.59815216 for the Tester service our algorithm predicted the price of 352.823. [Fig. 1]

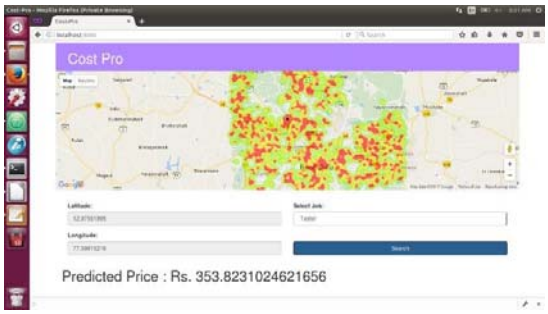

Fig 1. The web interface to show the predicted price

Our data in database for the nearby location with the coordinates 12.94380514, 77.58255235, for Tester service have the price of Rs. 343.9871184. So, in comparison with our prediction method the chances of error are very less.

Below will take a look at the results of time series forecasting using the graph

[Fig. 2] Depicts the loading of the price dataset with Pandas [7] function and dataset is baselined in an arbitrary year and month starting from 2000 to 2016, and plotted on graph.

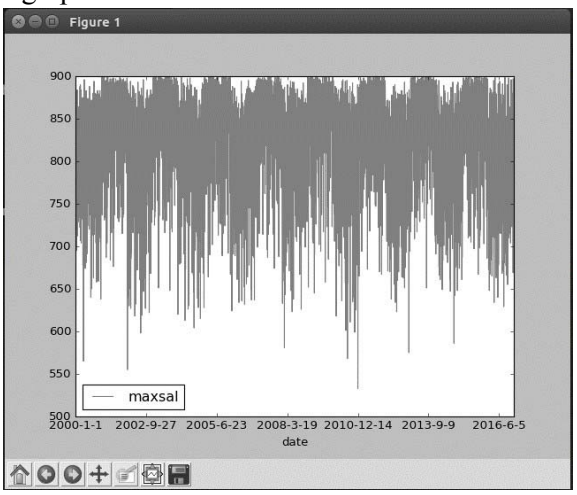

Fig2. Time series dataset plot

The autocorrelation plot of the time series is shown in [Fig.3]. This also built on Pandas [7].

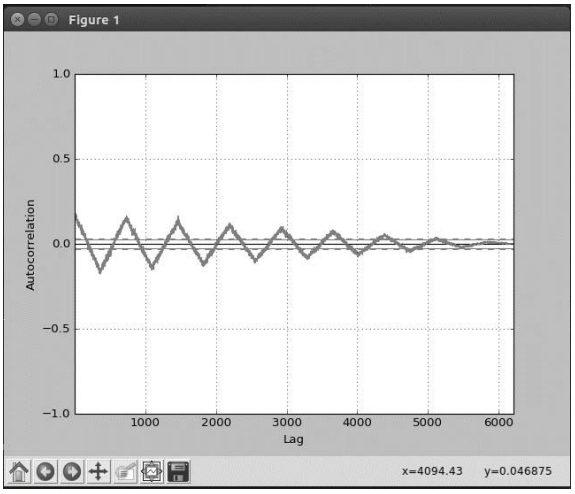

Fig3. Autocorrelation plot.

The output of the time series forecasting is shown in the figure [Fig. 4]. The line plot created showing the rolling forecast prediction of the service value in yearly manner with green in color.

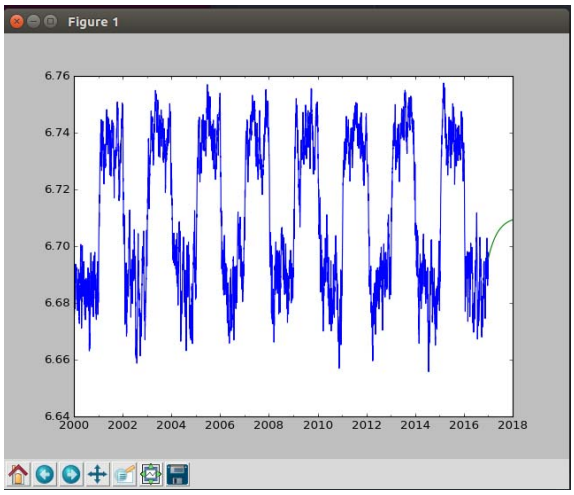

Fig 4: Predicted result for the time series data

VI. DISCUSSION AND CONCLUSIONS

1. Other possible machine learning techniques

Neural networks are most commonly used for the classification of task. However arbitrary function can be fitted easily. [5] Further efforts could be spent into applying neural network regression with plenty of samples.

2. Adding of more features

We can also add more features to the existing data set, and features may include like skillsets level, age groups, expertize level etc. Once the number feature increases its easier to predict the accurate value.

3. Algorithm

Linear regression is one of oldest and powerful way to predict the unknown value of a variable from the known value of another variable. And it also consists of finding the best fitting straight line through the points.

REFERENCE

1. The year of hyper local marketing https://balihoo.com/2017-emerging trend-hyper-local-marketing

2. Changes in relative wages, 1963-1987: Supply and demand factors. www.hi.is/~ajonsson/kennsla2003/relative _wages.pdf

3. Multiple regression http://www. stat.yale. edu/ Courses/1997-98 /101 /linmult.htm

4. How to create ARIMA model for time series forecasting in python http://machinelearningmastery.com/ari ma-for-time-series-forecasting-with- python/

5. T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. R¨uckstieß, and J. Schmidhuber, "Py- Brain," Journal of Machine Machine Learning Research, vol 11, pp.743-746, 2010

6. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

7. Wes McKinney "pandas: a Foundational Python Library for Data Analysis and Statistics"https://www.researchgate.net/publication/265194455_pandas_a_Foundational_Python_Library_for_Data_Analysis_an d_Statistics