Case Study - Pricing Forecast

Spring & River

Problem Statement:

Telecom job orders are being dispatched to the technicians who perform the job on premise of the customer. After completion, technician is moved to the next job allocated to the next available time window. A technician's time taken for the whole day is priced. The dataset tabulates various daily assignments given to the technicians.

Our goal is to predict the actual time taken by each technician at the job premise. This information would be processed in determining the price structure. The company already uses an adhoc method for determining the time by calculating the mean job completion time from the historical data. The challenge is to outperform the company benchmark and to put forward a more transparent and stable solution which will add to the business value.

Dataset:

The dataset includes JAN/FEB data for the year 2018. No of variables are **85** and total observations are **248965.** A data dictionary is maintained for proper understanding of the variables.

Data Processing and Feature engineering:

All observations with negative or missing values of job time have been removed. After understanding the business we settle down with 33 predictors from the original set of 85 variables. We plan to explore more with the feature engineering as the dimension of the problem is still very high. Variable selection is attained using various methods like Boruta, stepwise regression etc,. Boruta finds all features which are either strongly or weakly relevant to the target variable. The final set reduces to 20 variables.

Explore and prune Outliers:

Since almost all variables in the dataset are categorical in nature, predicting continuous response out of all nominal variables pose a complexity in achieving the goal. Outliers have been detected in two steps.

1. First we filter out job-time. The distribution plot records time as low as 1min or as high as 10 hrs or even more. After discussion, we select job-time in the range from 1hr to 6 hrs only. A qq-plot of job-time shows linear behavior in that range too.
2. Finding outliers from set of categorical variables is an ongoing research. Most outlier detection methods like cook's distance, works with continuous variable. We implement AVF algorithm which set a probability score against each row. Less score implies that particular observation has rare occurrence and we can drop it. We obtain AVF score in the range from 0.29 to 0.65. Cut off score is set to 0.31.

Exploratory Analysis:

Univariate analysis of the predictors has been done. The histogram of job-time (target) reveals near normal distribution. The bar plot of categorical variable reveals inhomogeneous distribution of corresponding levels for some predictors. The skewness issue is tackled by collapsing the lower levels into a new level named as "Other". In this way we don't lose any observations from the dataset. Average job-time vs. job type in ascending order found to be linear. This also indicates applicability of the linear regression model on the dataset.

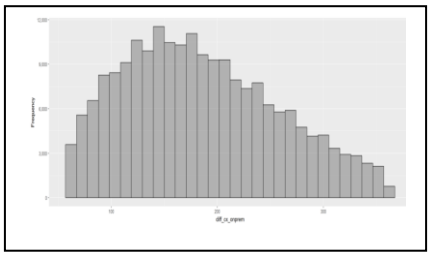
Sprinriver Technology Private Limited
Registered Office: Suite # C-2F, 302 A P C Road, KOLKATA – 700 009, India
Delivery Centre: EN-10, Touchstone Tower, Sector 5, Salt Lake, Kolkata – 700091, India

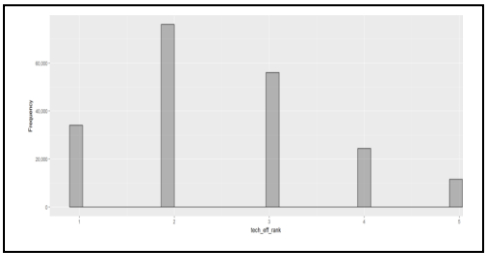
Train Model:

The clean data has been split into train and test with 70:30 ratio. We build a multi linear regression model. The summary of the regression result shows a RSE = 0.72 hrs and Adjusted R-squared: 0.35. The poor Adj. R- squared hints at 65% of the variability of target i.e., job-time is not explained by the predictors. Accuracy of prediction on the test data is found to be 60% in a 45 minutes error interval. If we increase the error interval to 90 minutes the accuracy improves to 85%. Low Adj. R-square implies that we need to re engineer features for better modeling.

Feature engineering:

It is observed that there is a variation in job-time against identical predictor levels. This variation can be attributed to human efficiency of the technicians. However, no such data is provided in the dataset. We devise a 5 point scoring system for all the technicians as learned from the historical data available. The following graph demonstrates the method.





The plot in the left is frequency distribution of job-time for a specific region and for a specific job category. The plot in right is the efficiency Factor derivation on 5-point scale for all technicians for the specific region and for the specific job category referred in Fig in left.

The score is considered as one of the predictors in the model. The regression model provides coefficients to be used in predicting the time. We assume scoring system will stabilize with large amount of data as it merely reflects efficiency of the technicians. Newly recruited technicians will not have any score as the model does not learn about their efficiency. In this case we can safely assign an average score (e.g., 3) and proceed with the prediction.

Discussion:

Comparison of the accuracy with efficiency and without has been shown in the chart below. We also include result for particular wire centre.

Sample Size	Train	Test	(+/- 30 min)	(+/- 45) min	(+/- 60) min	Adjusted R^2	RSE min	Active efficiency
202027	141420	60602	53%	71%	83%	0.62	44	Y
	141420	60602	41%	59%	73%	0.45	55	N

The result shows improve upon old model in terms of accuracy and Adjusted R-squared.
Several non linear model like SVR, GBM. GLM, CUBIST, Q-Reg, DNN was also applied on the dataset. No significant improvement was noticed.

Conclusion:

Pricing model for telecom job orders have been discussed in machine learning frame work. We observe that current dataset is not yielding the desired result. The linear model shows low Adjusted R-squared value. However, improvement was achieved by introducing new features like efficiency of the technicians. For robust result, careful estimation of job-time is very essential.

As a next set of assignment, GPS data is to be used with an expectation to improve the prediction window. Other advanced method like xgboost with parameter tuning can also boost the accuracy.

Sprinriver Technology Private Limited
Registered Office: Suite # C-2F, 302 A P C Road, KOLKATA – 700 009, India
Delivery Centre: EN-10, Touchstone Tower, Sector 5, Salt Lake, Kolkata – 700091, India