Data Mining

Zomato Restaurant Reviews

The food industry is one of the most important industries in the world. It is important for restaurants to have quality food and good service. While most of the time reviews of a good restaurant goes out to people through people, most of them rely on online ratings and reviews. The current trend is to google and find out about anything and everything. Hence it is important that restaurants have a very good rating. Hence, the project is aimed at how a restaurant should increase their rating in order to appear among the top.

Things like the location, the cuisine, the service, the quality and many more are important for a good restaurant. But there are many more extra things that can be a helping hand to a better restaurant. The project is done on the restaurant reviews from Zomato which was published in Kaggle.

CRISP DM methodology is followed in this project as it is the best model compared to the other data mining methodology such as SEMMA and KDD. The abbreviation for CRISP DM is Cross-Industry standard process for Data Mining. CRISP DM has several stages such as Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. This gives a structed methodology for implementing a data mining project.

Business Problem

The dataset will be analyzed to help a business set up a new restaurant. The attributes such as the “average cost for two in dollars”, “has table booking” and “has online delivery” will be considered to analyze if there is an effect on the average rating. In other words, the aim is to see if the mentioned three attributes are important to get good reviews from customers. When a restaurant gets good rating, automatically the business would increase. The stakeholders will be the management.

Analytics Problem

The analytical problem is to see if the independent attributes “average cost for two in dollars”, “has table booking” and “has online delivery” is significant with the dependent variable “average rating”.

Data Source

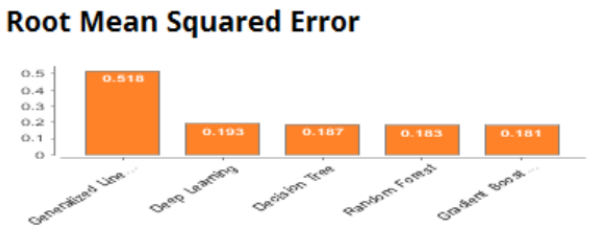
The data for the project has been taken from Kaggle, which is an online open data source that has plenty of different data. The dataset that has been chosen has more than 9000 rows and plenty of attributes such as country, rate for two, longitude, latitude and many more. Some of the rows were removed as it did not have the names of the restaurants. While few others did not have the average price.

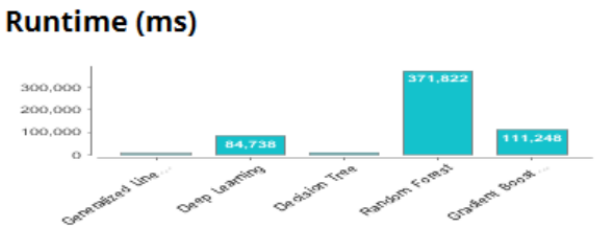
The restaurants are from different places over the world, hence the average cost for two has been converted to dollars from the different currencies available.

Methodology Selection

To establish a relationship between dependent and independent attributes, Linear regression was the choice. However, the dataset was run with the help of auto model and the results are show below. It was clear from the results that GLM or Linear Regression was the model to have a better performance.

Comparison in Rapid Miner





Model	Root Mean Squared Error	Runtime
Generalized Linear Model	0.518	5 s
Deep Learning	0.193	1 min 24 s
Decision Tree	0.187	8 s

Model	Root Mean Squared Error	Runtime
Decision Tree	0.187	8 s
Random Forest	0.183	6 min 11 s
Gradient Boosted Trees	0.181	1 min 51 s

Model Building

The dataset is separated into two train set and test set. 80% of the data is taken as the train set and 20% is taken as a test set. Simple linear regression and multi linear regression methods are used in this project.

These are the results of simple linear regression applied on the independent attributes selected.

```
> summary(a_train)

Call:
lm(formula = Aggregate.rating ~ Avg.cost.for.two.in.Dollars,
    data = trainset)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6225  -0.4703   0.5940   0.9828   2.5720

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.261301   0.020814  108.64   <2e-16 ***
Avg.cost.for.two.in.Dollars 0.029658   0.001086   27.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.447 on 7204 degrees of freedom
  (386 observations deleted due to missingness)
Multiple R-squared:  0.0938,    Adjusted R-squared:  0.09368
F-statistic: 745.7 on 1 and 7204 DF,  p-value: < 2.2e-16
```

```
> summary(a_test)

Call:
lm(formula = Aggregate.rating ~ Avg.cost.for.two.in.Dollars,
    data = testset)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7876 -0.4866  0.6085  0.9908  2.4908

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.285786   0.040837   55.97   <2e-16 ***
Avg.cost.for.two.in.Dollars 0.027431   0.001994   13.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.448 on 1816 degrees of freedom
  (81 observations deleted due to missingness)
Multiple R-squared:  0.09442,   Adjusted R-squared:  0.09392
F-statistic: 189.3 on 1 and 1816 DF,  p-value: < 2.2e-16
```

```
> summary(a_train)

Call:
lm(formula = Aggregate.rating ~ Has.Table.booking, data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4270 -0.3270  0.5461  1.0461  2.3461

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.55390    0.01822  140.15   <2e-16 ***
Has.Table.bookingYes  0.87309    0.05281   16.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.49 on 7590 degrees of freedom
Multiple R-squared:  0.03476,   Adjusted R-squared:  0.03464
F-statistic: 273.3 on 1 and 7590 DF,  p-value: < 2.2e-16
```

```
> summary(a_test)

Call:
lm(formula = Aggregate.rating ~ Has.Table.booking, data = testset)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4921 -0.3921  0.4725  0.9725  2.3725

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.52752    0.03653  69.199   <2e-16 ***
Has.Table.bookingYes 0.96457    0.10007   9.639   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.482 on 1897 degrees of freedom
Multiple R-squared:  0.04669,   Adjusted R-squared:  0.04619
F-statistic: 92.91 on 1 and 1897 DF,  p-value: < 2.2e-16


> summary(a_train)

Call:
lm(formula = Aggregate.rating ~ Has.Online.delivery, data = trainset)

Residuals:
   Min     1Q Median     3Q    Max
-3.235 -0.635  0.465  1.041  2.441

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.45894    0.01968  124.97   <2e-16 ***
Has.Online.deliveryYes 0.77605    0.03886   19.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.478 on 7590 degrees of freedom
Multiple R-squared:  0.04991,   Adjusted R-squared:  0.04979
F-statistic: 398.7 on 1 and 7590 DF,  p-value: < 2.2e-16

> summary(a_test)

Call:
lm(formula = Aggregate.rating ~ Has.Online.delivery, data = testset)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3002 -0.7002  0.4767  0.9998  2.4767

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.42330    0.03929   61.68   <2e-16 ***
Has.Online.deliveryYes 0.87690    0.07627   11.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.467 on 1897 degrees of freedom
Multiple R-squared:  0.06515,   Adjusted R-squared:  0.06466
F-statistic: 132.2 on 1 and 1897 DF,  p-value: < 2.2e-16
```

These are the results of multi linear regression applied on the independent attributes selected.

```
> summary(a_train)

Call:
lm(formula = Aggregate.rating ~ Avg.cost.for.two.in.Dollars +
    Has.Table.booking + Has.Online.delivery, data = trainset)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2762  -0.9061   0.4899   1.0139   2.7953

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.002944   0.022386  89.475  < 2e-16 ***
Avg.cost.for.two.in.Dollars 0.027138   0.001124  24.141  < 2e-16 ***
Has.Table.bookingYes         0.379083   0.053431   7.095 1.42e-12 ***
Has.Online.deliveryYes       0.883233   0.036925  23.919  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.385 on 7202 degrees of freedom
  (386 observations deleted due to missingness)
Multiple R-squared:    0.17,     Adjusted R-squared:  0.1696
F-statistic: 491.6 on 3 and 7202 DF,  p-value: < 2.2e-16
> summary(a_test)

Call:
lm(formula = Aggregate.rating ~ Avg.cost.for.two.in.Dollars +
    Has.Table.booking + Has.Online.delivery, data = testset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8351 -0.8910  0.4436  1.0171  2.7906

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.999271   0.043770  45.677  < 2e-16 ***
Avg.cost.for.two.in.Dollars 0.024483   0.002067  11.847  < 2e-16 ***
Has.Table.bookingYes         0.411417   0.102458   4.015 6.17e-05 ***
Has.Online.deliveryYes       0.947917   0.072673  13.044  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.374 on 1814 degrees of freedom
  (81 observations deleted due to missingness)
Multiple R-squared:  0.1862,    Adjusted R-squared:  0.1848
F-statistic: 138.3 on 3 and 1814 DF,  p-value: < 2.2e-16
```

Our analytical question is 'is there a relationship among "average cost for two in dollars", "has table booking", "has online delivery" and "Average rating"?' This is answered by fitting the model multi linear regression by testing the hypothesis.

The F-statistic can be used to determine whether we should reject this null hypothesis. In this case the p-value corresponding to the F-statistic is very low, indicating clear evidence of a relationship between "average cost for two in dollars", "has table booking" and "has online delivery" and "Average rating".

How strong is the relationship?

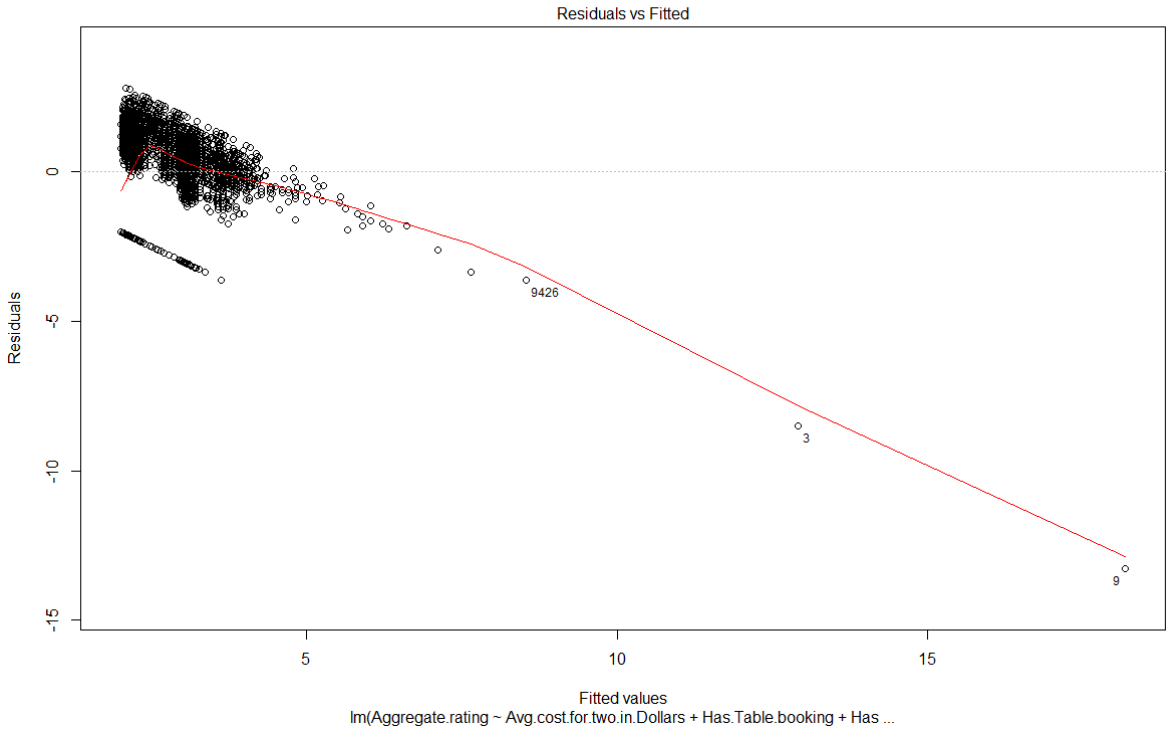
Consider two measures of model accuracy. First, the RSE estimates the standard deviation of the response from the population regression line. Second, the $R2$ statistic records the percentage of variability in the response that is explained by the predictors.

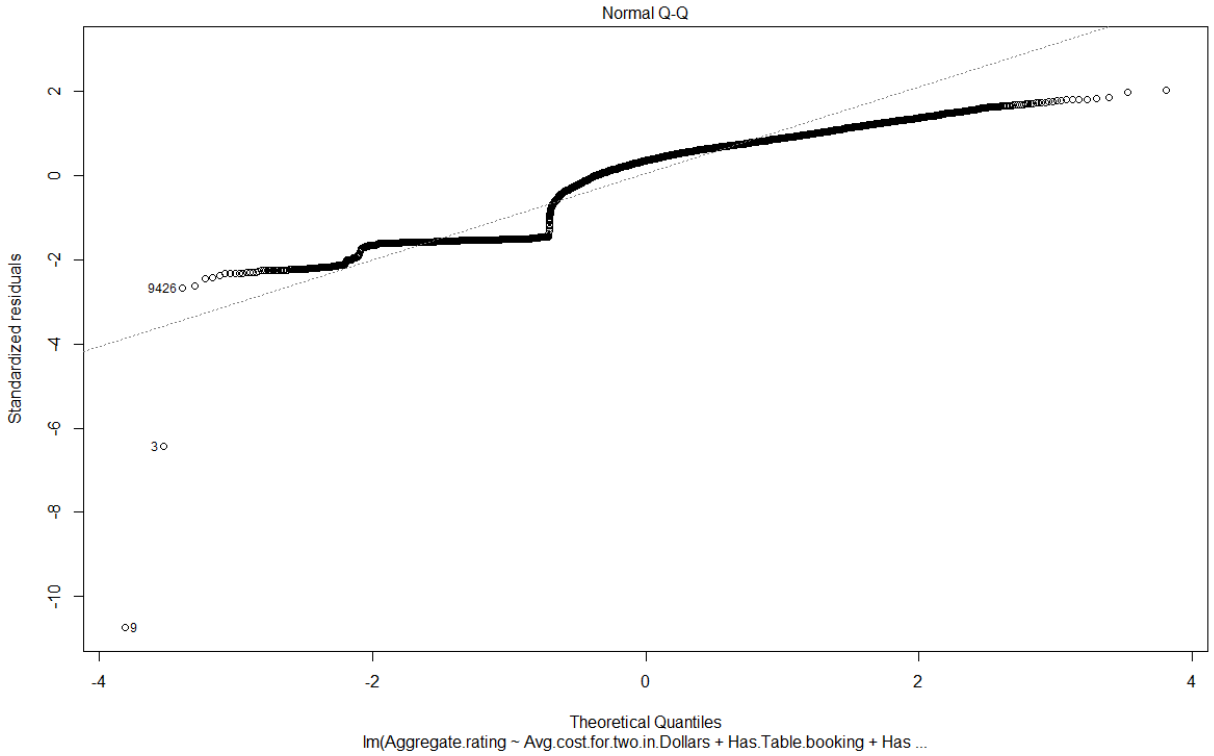
How can we predict the relationship accurately?

$y = 2.002944 + 0.027138 * Avg.cost.for.two.in.Dollars + 0.379083 * Has.Table.booking + 0.883233 * Has.Online.delivery$

The accuracy associated with this estimate depends on whether we wish to predict an individual response, $y=f(x)+\varepsilon$, or the average response, $f(x)$. If the former, we use a prediction interval, and if the latter, we use a confidence interval. Prediction intervals will always be wider than confidence intervals because they account for the uncertainty associated with ε, the irreducible error. Prediction interval = $\hat{y}\pm 2 * residual\ standard\ error$.

(Note: For the model to be useful the four assumptions of the error should approx. hold:
1. Normally distributed
2. Mean of 0
3. Equal Variance
4. Probabilistically independent of residual error of previous term
A transformation (e.g. log; sqrt) can be applied to the predictors to normalize the residual error).





Residual plots can be used to identify non-linearity. If the relationships are linear, then the residual plots should display no pattern. The inclusion of transformations of the predictors in the linear regression model can be made to accommodate non-linear relationships.

To conclude, we see a significant relationship with table bookings, average cost for two and online delivery on the average rating of restaurants.

Deployment

This can be deployed as an online link for restaurants to ensure that they have considered all the important things to open a restaurant and on the long run get good ratings thus increasing their business in the future.