**Final Project Report**

**A project on**

**Facebook Comment Volume Dataset**

**By**

**Group 8:**

**Arunabho Chattopadhyay**

**Deepesh Jain**

**Khyati Mishra**

**Manasi Vivek Kher**

**Tejashree Pandit**

**1. Executive Summary:**

The objective of this project is to create a predictive model that can estimate the number of comments a Facebook post receives. Businesses can get insights on the length of post, effective time to publish the post so that they get majority attention from consumers. This would be helpful for them to decide on their marketing strategies on social media platform. Similarly, any individual other than businesses would get an overview of how parameters like length, time of post, popularity of page can make an impact on the number of comments the post receives.

Our aim is to model the user comment patterns over the posts on Facebook Pages and predict on an average how many comments a post is expected to receive in next H hrs. This can be achieved by applying various modelling techniques like Regression, Decision Trees, and neural networks on different data-set variants.

In this report, we present to you our findings using the data mining techniques mentioned above.

**2. Background:**

The amount of data that is uploaded to social networking services is increasing day by day. So, there is a massive requirement to study the highly dynamic behavior of users towards these services. More than 1.8 billion people use Facebook every month to connect with friends and family and to discover things that matter. Marketing on Facebook helps businesses find new customers and build lasting relationships with them. Social media is increasingly being used as a marketing tool. This analysis will be helpful to businesses to better understand the consumer behavior.

**3. Data Exploration and Cleaning:**

Detailed description of the dataset variables is as follows:

|  |  |  |
| --- | --- | --- |
| **No.** | **Variable** | **Description** |
| 1 | Page Popularity/likes | Defines the popularity or support for the source of the document |
| 2 | Page Checkins | Describes how many individuals so far visited this place. This feature is only associated with the places eg: some institution, place, theater etc. |
| 3 | Page talking about | Defines the daily interest of individuals towards source of the document/ Post. The people who actually come back to the page, after liking the page. This include activities such as comments, likes to a post, shares, etc by visitors to the page |
| 4 | Page Category | Defines the category of the source of the document eg: place, institution, brand etc. |
| 5 - 29 | Derived | These features are aggregated by page, by calculating min, max, average, median and standard deviation of essential features |
| 30 | CC1 | The total number of comments before selected base date/time |
| 31 | CC2 | The number of comments in last 24 hours, relative to base date/time |
| 32 | CC3 | The number of comments in last 48 to last 24 hours relative to base date/time |
| 33 | CC4 | The number of comments in the first 24 hours after the publication of post but before base date/time |
| 34 | CC5 | The difference between CC2 and CC3 |
| 35 | Base Time | Selected time in order to simulate the scenario |
| 36 | Post length | Character count in the post |
| 37 | Post Share Count | This features counts the no of shares of the post, that how many peoples had shared this post on to their timeline |
| 38 | Post Promotion Status | To reach more people with posts in News Feed, individual promote their post and this features tells that whether the post is promoted (1) or not(0) |
| 39 | H Local | This describes the H hrs, for which we have the target variable/ comments received |
| 40-46 | Post published weekday | This represents the day(Sunday...Saturday) on which the post was published |
| 47-53 | Base DateTime weekday | This represents the day (Sunday...Saturday) on selected base Date/Time |
| 54 | Target | The no of comments in next H hrs (H is given in Feature no 39) |

Data Exploration and Cleaning:

* Exploration: The data has over 40k rows 53. The target/dependent variable is ‘No of comments’ which indicates the total number of comments received in H hours. We made use of graph explore to identify whether there is any missing data present in the dataset.
* Cleaning:

To identify the significant variables, we performed regression on the entire dataset.

Following are the results of regression:

It can be seen from the results that there are 10 significant variables. We constructed box plots to find the outliers in significant variables.

The sheet containing box plots and outliers’ analysis is attached here.

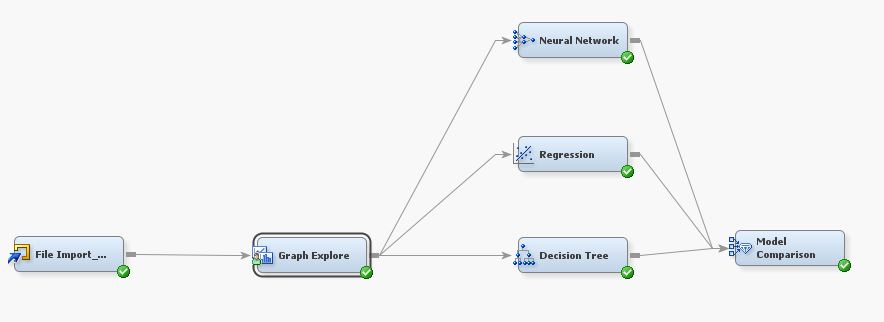


Before proceeding further, we removed the outliers from some of the significant variables as stated in the excel sheet attached above.

**4. BI Models and Techniques:**

We have predicted using following data mining techniques:

**STEP 1: Identifying the best method**

****

* **Multiple Linear Regression:**

We ran regression on the entire dataset to determine the significant variables

The results are as follows:

Model Fit Statistics

R-Square 0.3252 Adj R-Sq 0.3244

AIC 276312.8892 BIC 276315.0066

SBC 276735.2732 C(p) 49.0000

Analysis of Maximum Likelihood Estimates

Standard

Parameter DF Estimate Error t Value Pr > |t|

Intercept 1 -3.7282 2.0250 -1.84 0.0656

Base\_DateTime\_weekday\_1 1 0.1114 0.5632 0.20 0.8433

Base\_DateTime\_weekday\_2 1 1.1634 0.6199 1.88 0.0606

Base\_DateTime\_weekday\_3 1 0.5220 0.6553 0.80 0.4257

Base\_DateTime\_weekday\_4 1 1.4907 0.6466 2.31 0.0211

Base\_DateTime\_weekday\_5 1 0.4801 0.6091 0.79 0.4306

Base\_DateTime\_weekday\_6 1 0.5957 0.5659 1.05 0.2926

Base\_DateTime\_weekday\_7 0 0 . . .

Base\_time 1 -0.1924 0.00741 -25.97 <.0001

CC1 1 0.0442 0.0164 2.70 0.0070

CC2 1 0.2197 0.00373 58.96 <.0001

CC3 1 -0.0226 0.00351 -6.44 <.0001

CC4 1 -0.0683 0.0183 -3.74 0.0002

CC5 0 0 . . .

Derived\_1 1 -0.5684 0.1219 -4.66 <.0001

Derived\_10 1 -0.0324 0.0399 -0.81 0.4173

Derived\_11 1 0.1116 0.1140 0.98 0.3276

Derived\_12 1 0.0120 0.00402 2.99 0.0028

Derived\_13 1 0.3216 0.0782 4.11 <.0001

Derived\_14 1 -0.0250 0.0402 -0.62 0.5351

Derived\_15 1 -0.1406 0.0352 -4.00 <.0001

Derived\_16 1 0.5264 0.1342 3.92 <.0001

Derived\_17 1 -0.0164 0.00652 -2.52 0.0117

Derived\_18 1 -1.0464 0.1262 -8.29 <.0001

Derived\_19 1 -0.1495 0.0588 -2.54 0.0111

Derived\_2 1 0.0144 0.00621 2.32 0.0204

Derived\_20 1 0.2144 0.0873 2.45 0.0141

Derived\_21 1 0.000930 0.00372 0.25 0.8026

Derived\_22 1 0.00219 0.00536 0.41 0.6827

Derived\_23 1 0.1971 0.0555 3.55 0.0004

Derived\_24 1 -0.0323 0.0268 -1.20 0.2291

Derived\_25 1 0.00973 0.0239 0.41 0.6844

Derived\_3 1 0.9640 0.1225 7.87 <.0001

Derived\_4 1 0.1328 0.0587 2.26 0.0236

Derived\_5 1 -0.1608 0.0827 -1.94 0.0519

Derived\_6 1 -0.1170 0.0416 -2.81 0.0050

Derived\_7 1 -0.00375 0.00590 -0.64 0.5249

Derived\_8 0 0 . . .

Derived\_9 1 0.1307 0.0332 3.93 <.0001

H\_Local 1 0.4130 0.0788 5.24 <.0001

Page\_Category 1 -0.0101 0.00769 -1.32 0.1876

Page\_Checkins 1 -0.00001 7.571E-6 -1.67 0.0940

Page\_Popularity\_likes 1 2.323E-8 2.866E-8 0.81 0.4178

Page\_talking\_about 1 -0.00002 2.271E-6 -8.50 <.0001

Post\_Length 1 -0.00003 0.000385 -0.09 0.9322

Post\_Promotion\_Status 0 0 . . .

Post\_Share\_count 1 0.00291 0.000177 16.46 <.0001

Post\_published\_weekday\_1 1 -0.0107 0.5926 -0.02 0.9856

Post\_published\_weekday\_2 1 -0.1522 0.6238 -0.24 0.8072

Post\_published\_weekday\_3 1 -0.0486 0.6546 -0.07 0.9408

Post\_published\_weekday\_4 1 1.1798 0.6513 1.81 0.0701

Post\_published\_weekday\_5 1 0.8018 0.6184 1.30 0.1948

Post\_published\_weekday\_6 1 1.0644 0.5673 1.88 0.0606

Post\_published\_weekday\_7 0 0 . . .

We then rejected the least significant variables and ran regression again.

The results are as follows:

Model Fit Statistics

R-Square 0.3236 Adj R-Sq 0.3234

AIC 276346.0868 BIC 276348.1009

SBC 276492.6282 C(p) 17.0000

Analysis of Maximum Likelihood Estimates

Standard

Parameter DF Estimate Error t Value Pr >|t|

Intercept 1 -2.2410 1.8459 -1.21 0.2247

Base\_time 1 -0.1928 0.00739 -26.10 <.0001

CC2 1 0.2154 0.00338 63.71 <.0001

CC3 1 -0.0227 0.00351 -6.45 <.0001

CC4 1 -0.0195 0.00288 -6.77 <.0001

Derived\_1 1 -0.2494 0.0917 -2.72 0.0065

Derived\_12 1 0.00764 0.00136 5.62 <.0001

Derived\_13 1 0.2878 0.0494 5.82 <.0001

Derived\_15 1 -0.0793 0.0177 -4.48 <.0001

Derived\_16 1 0.1480 0.1026 1.44 0.1492

Derived\_18 1 -1.0904 0.0599 -18.19 <.0001

Derived\_23 1 0.1719 0.0245 7.02 <.0001

Derived\_3 1 1.0098 0.0556 18.15 <.0001

Derived\_9 1 0.0846 0.0201 4.21 <.0001

H\_Local 1 0.3837 0.0760 5.05 <.0001

Page\_talking\_about 1 -0.00002 1.758E-6 -10.20 <.0001

Post\_Share\_count 1 0.00291 0.000176 16.56 <.0001

After removing the outliers and insignificant variables, we ran the regression and came up with the following results

Model Fit Statistics

R-Square 0.3309 Adj R-Sq 0.3307

AIC 275583.6381 BIC 275585.6440

SBC 275678.4561 C(p) 11.0000

Analysis of Maximum Likelihood Estimates

Standard

Parameter DF Estimate Error t Value Pr > |t|

Intercept 1 -2.0830 1.8281 -1.14 0.2545

Base\_time 1 -0.1833 0.00734 -24.98 <.0001

CC2 1 0.2258 0.00338 66.81 <.0001

CC3 1 -0.0199 0.00350 -5.68 <.0001

CC4 1 -0.0341 0.00291 -11.71 <.0001

Derived\_18 1 -0.5161 0.0407 -12.67 <.0001

Derived\_3 1 0.5311 0.0383 13.85 <.0001

Derived\_9 1 0.1513 0.0130 11.60 <.0001

H\_Local 1 0.3786 0.0752 5.03 <.0001

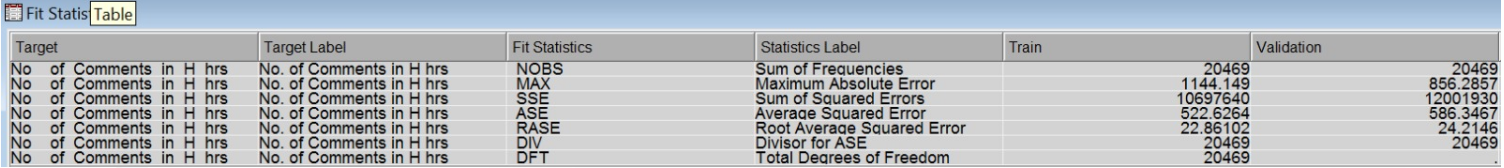
Page\_talking\_about 1 -0.00001 1.699E-6 -8.42 <.0001

Post\_Share\_count 1 0.0103 0.000343 29.96 <.0001

From the R-Sq value, we, therefore concluded that the model improves after preprocessing the data and rejecting the insignificant variables.

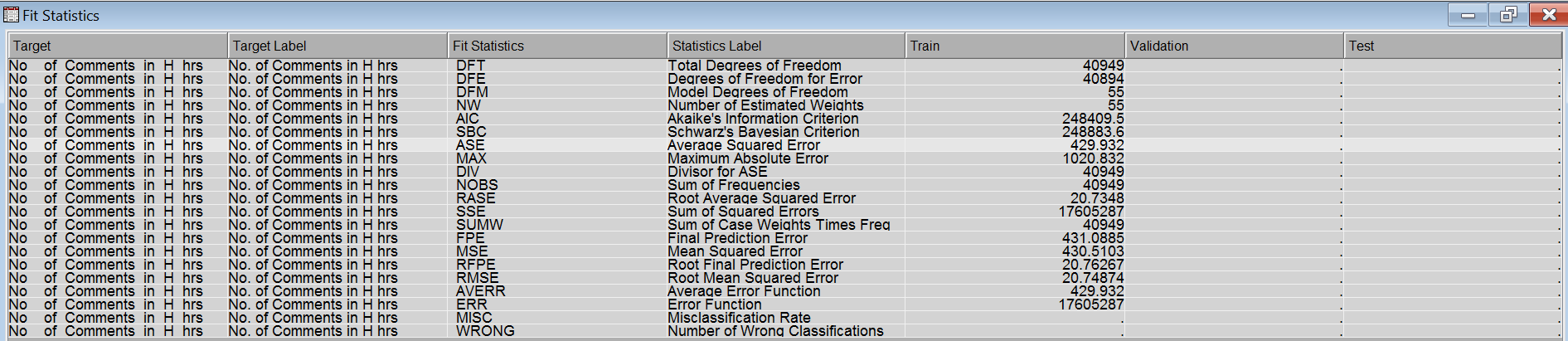
* **Decision Trees**

We then proceeded to perform the decision tree analysis on the processed data. The Fit Statistics of decision tree are as follows:

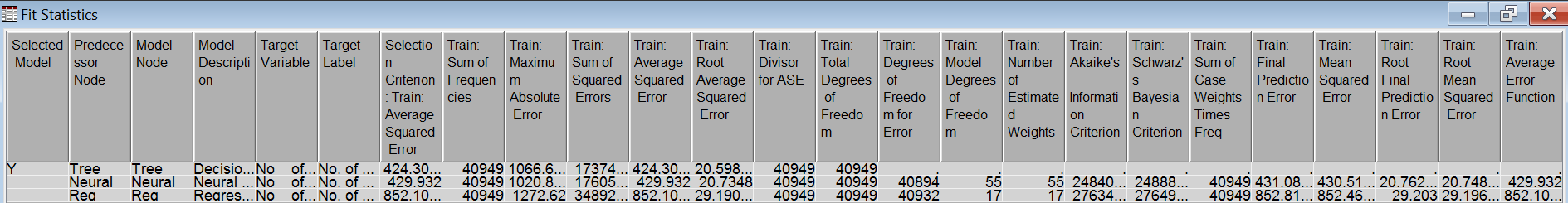


* **Neural networks:**

Other modelling techniques like neural networks was used to figure out the best fit model. The following are the fit statistics:

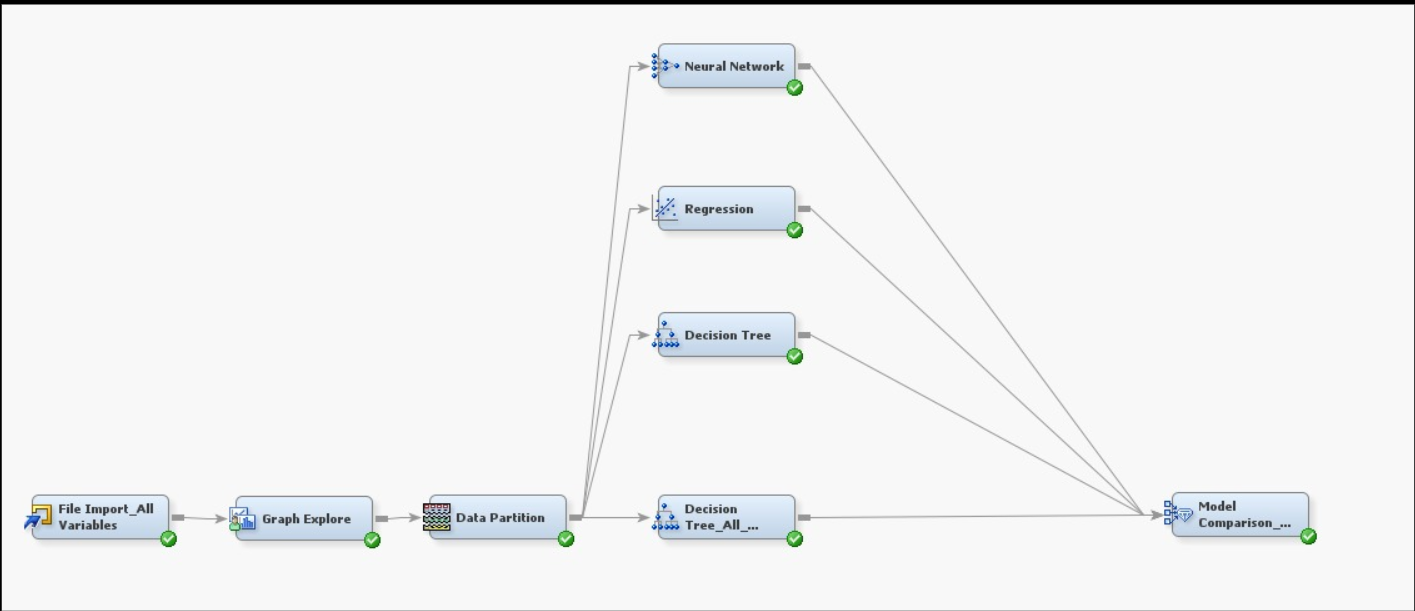


After obtaining regression, neural network and decision tree model, we used a model comparator to find the best fit model. The results are as follows:

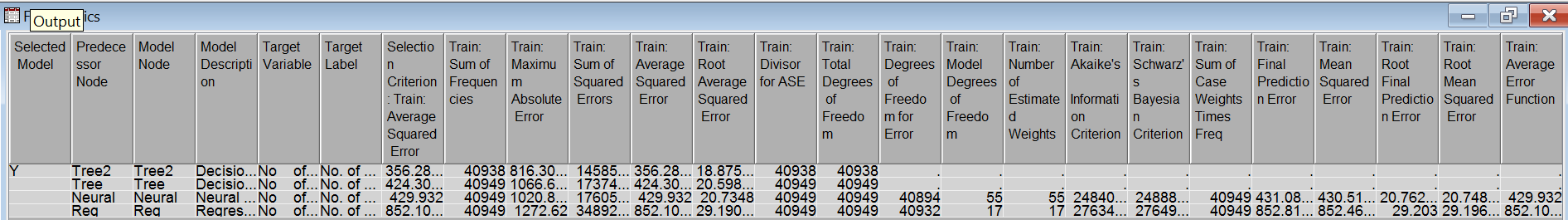


Since, the average squared error for decision tree is less than that of regression and neural networks, it is the better approach of the two models. On concluding that decision tree is a better model of all, we optimized the input variables used in the next steps.

**STEP 2: Improving the decision tree**

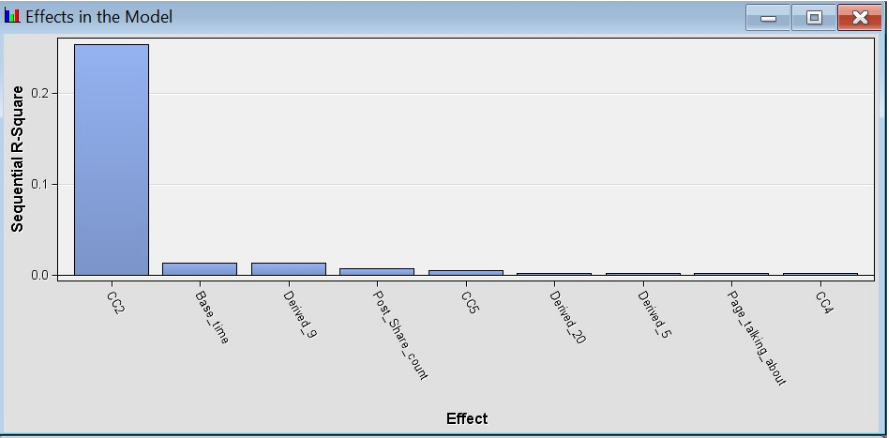


To further enhance the decision tree model, the input variables used were analyzed. Another decision tree was created whose input was all variables. It was compared with other models who had only significant variables as input. On comparing it with other models, the decision tree with all variables as input produced better results i.e had low average squared error. The following is the output of fit statistics:

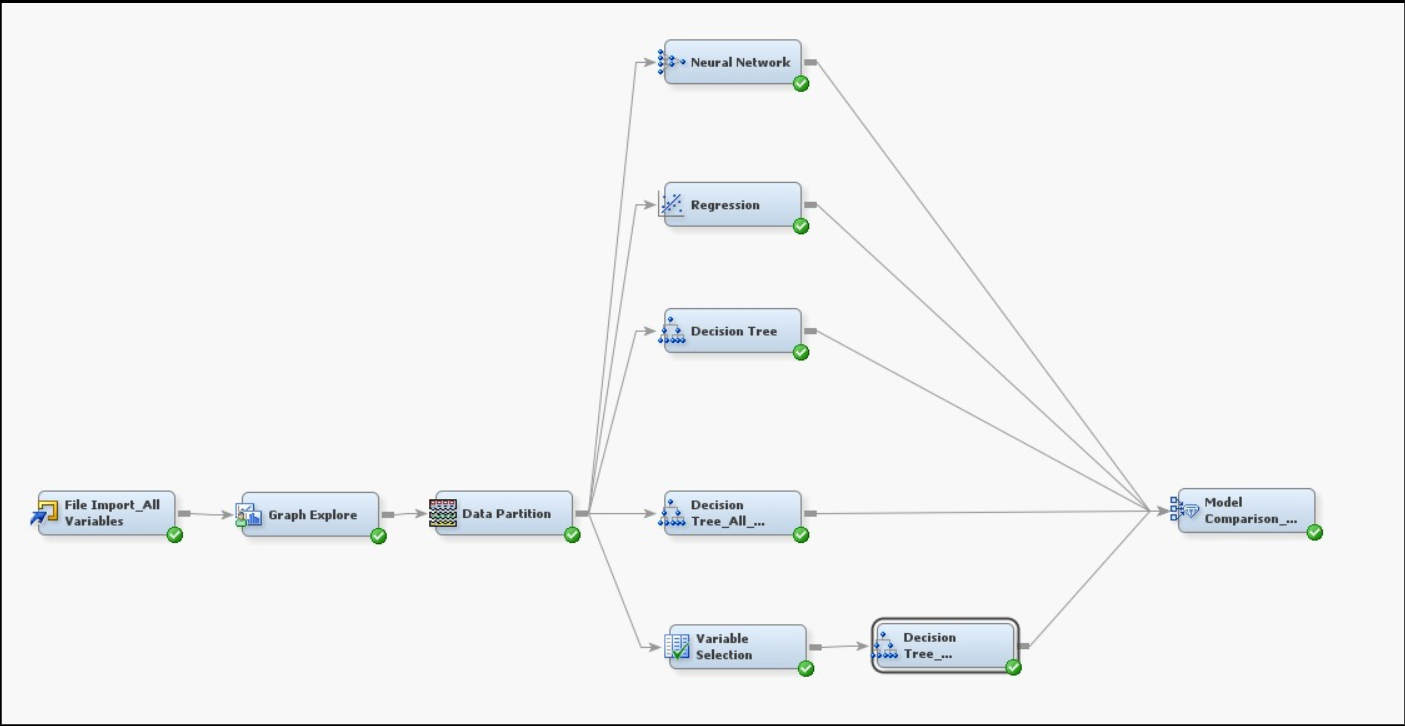


**STEP 3: Using Variable Selection(R-square) to improve the model**

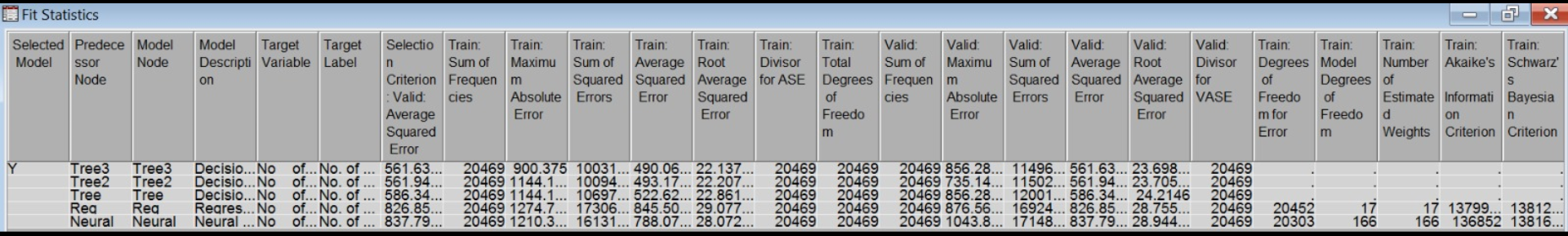
In order to optimize the model, variables were selected using R-square method of variable selection node.



A decision tree was created using only significant variables as input. It was then compared with previous decision tree (all variables as input).



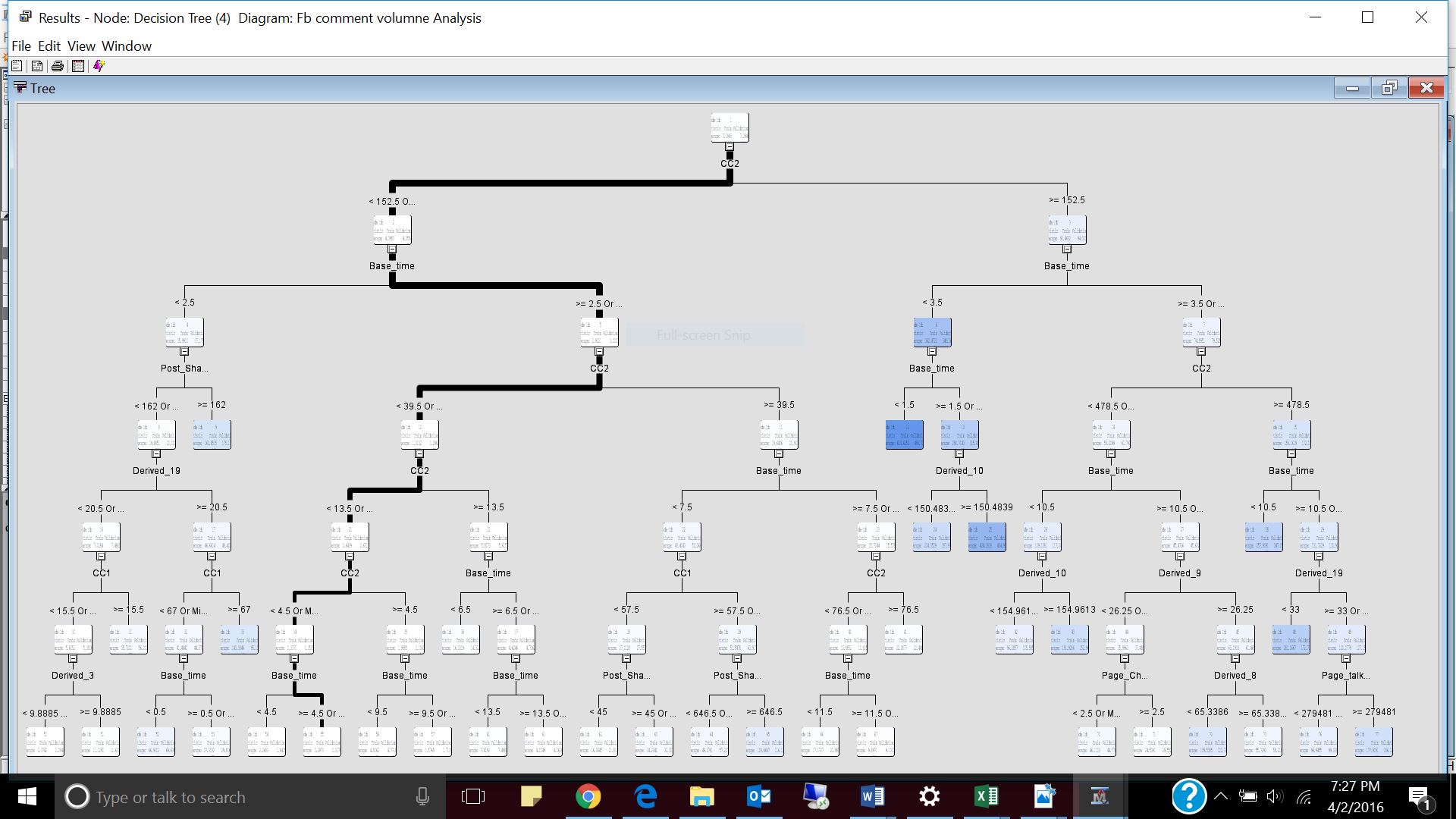
On analyzing the results, it was concluded that decision tree that was obtained using R-square technique was a better model as compared to the rest of the models. Following is the result of model comparator:



Decision tree3 (with variables selected by R-square) was a better model.

**5. Conclusion**

Interpretation of decision tree:



-The variable used in the primary node was CC2- the no of comments in last 24 hours relative to the base time.

- If the number of comments are less than 152.5 or missing then the average count for comments in 4.354 for validation data.

- Also, if the base time is >2.5hrs/missing then the average count of comments is 3 for validation data

- Likewise, CC2 and base time variables are considered predominantly in the decision tree

- Depending on the base time and CC2 variables, we can find out the average number of comments a post will receive with the help of a decision tree.

The overall final model for interpretation is as follows:

