**Business Problem**

Large number of apps are being developed everyday. As per Google there are about 2.8 million apps in Google playstore and most of them don’t make profit. We designed a model which can predict how much profit a new app can make in the market based on the number of installations of the app which will be helpful for the app developers.

**Dataset Description**

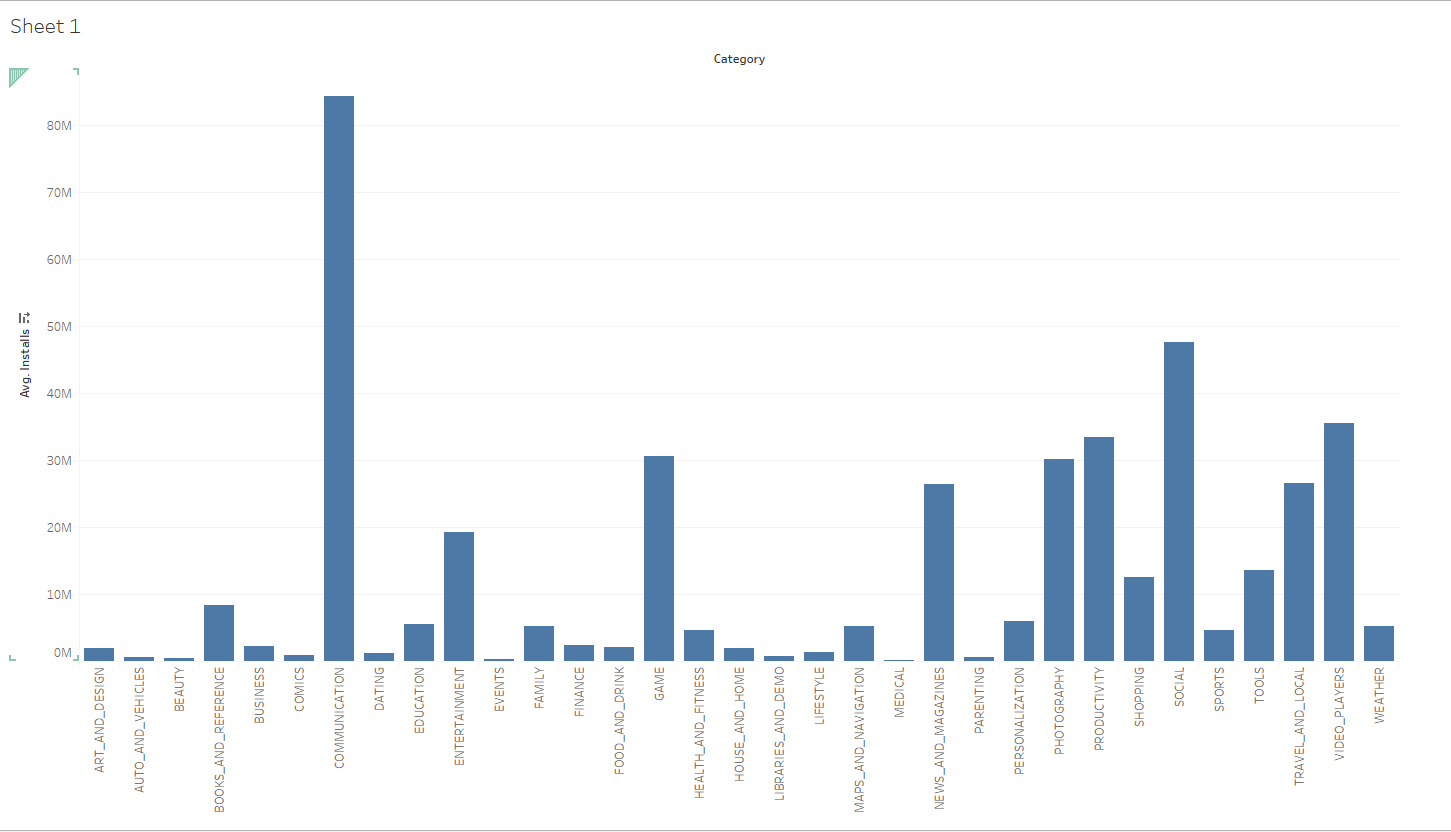
For our prediction we took Google Play Store dataset from online Community used by data scientist and Machine learners called kaggle.com. We have 12 columns and 10841 rows.

**The list of variables are**

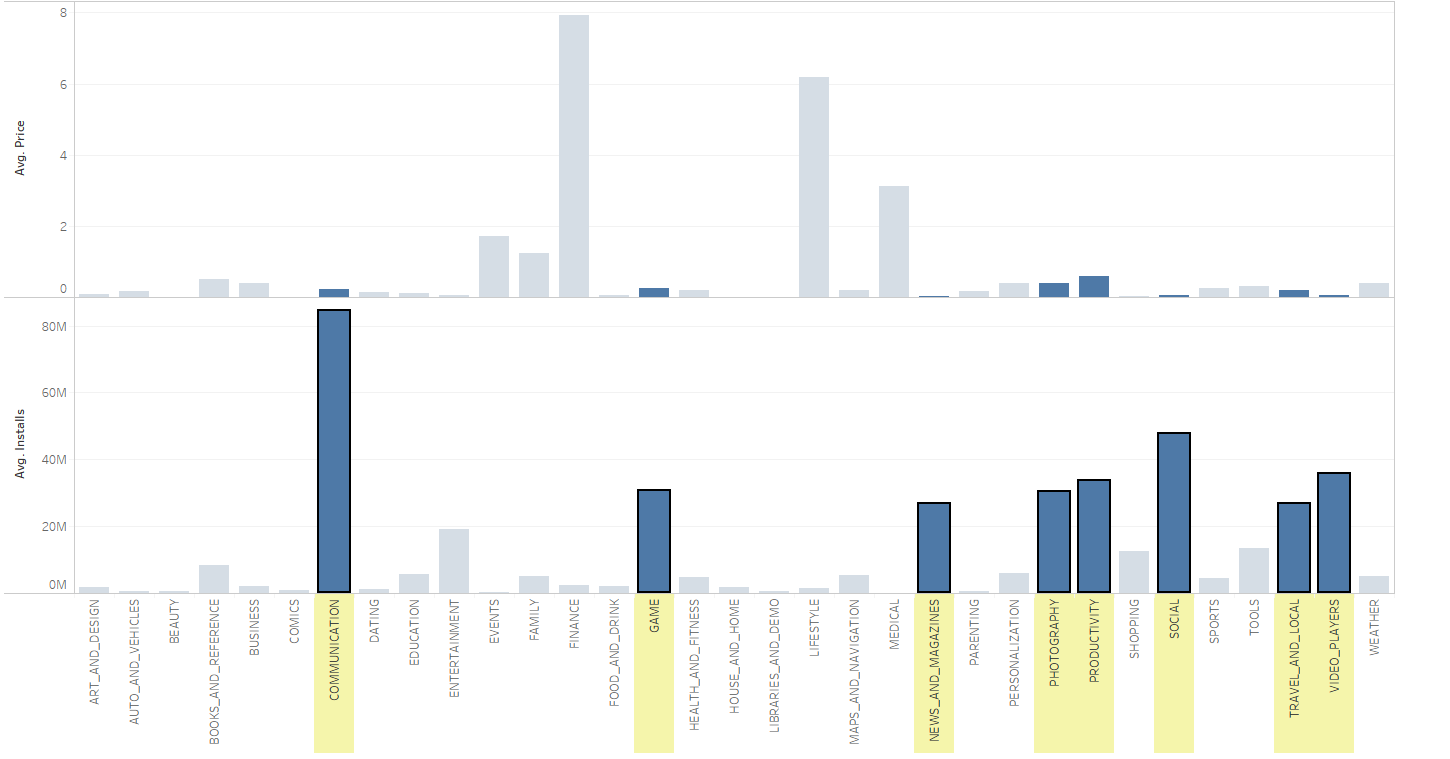
1. Category
2. Rating
3. Reviews
4. Size
5. Installs
6. Type
7. Price
8. Content Rating
9. Genres
10. Last Updated
11. Current Version
12. Android Version

We have **Installs** as **Target Variable**. The independent variables can be divided into variables developers have control on: **Size, Price & Type** and variables developers don’t have control on: **Reviews, Rating and Category**.

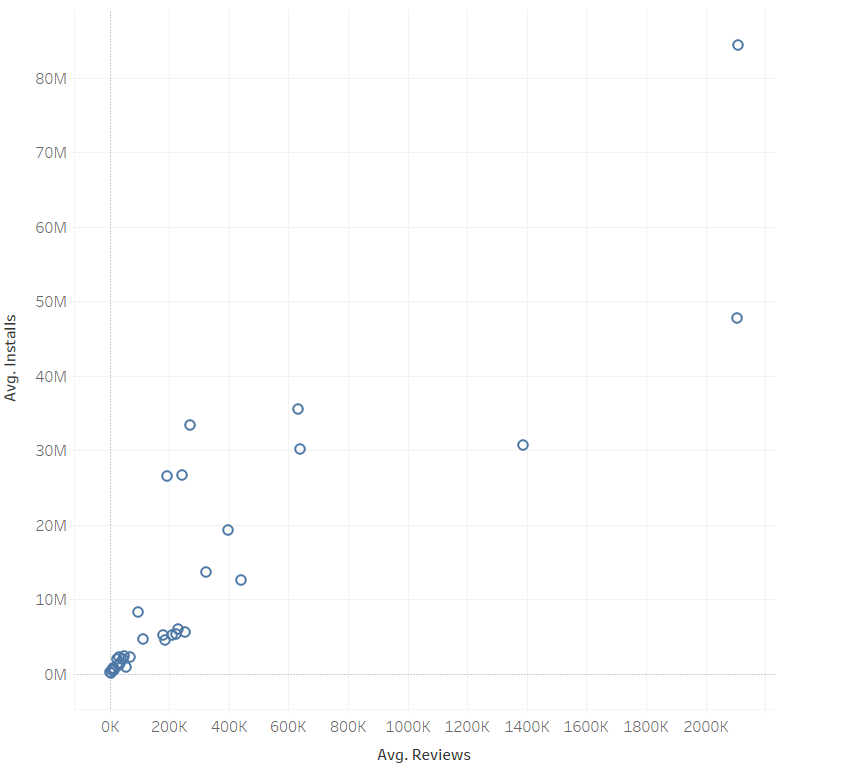
Data Visualization:



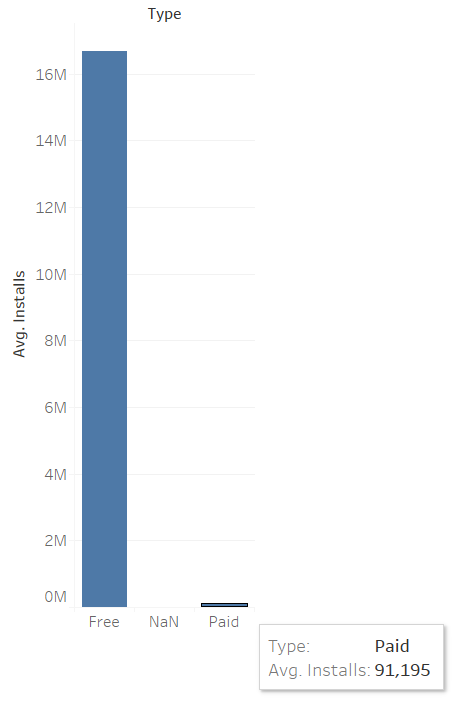
From the above graph its visible that the highest amount of installation are done for the applications that lie in the category of communications and second in the category of social.



From the graph plotted above i.e Price vs Installs we can clearly see how the price of the applications affects the number of installation. The Lower the pricing of the data the higher the number of installations.



From the above scatterplot its visible that data with more amount of reviews have been installed more amount of times.

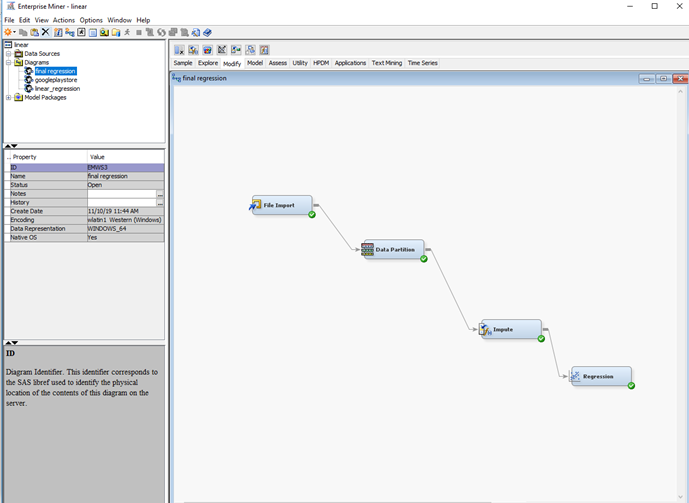


From the graph above we can infer that apps which require some form of payment are downloaded less compared to apps that are free to use.

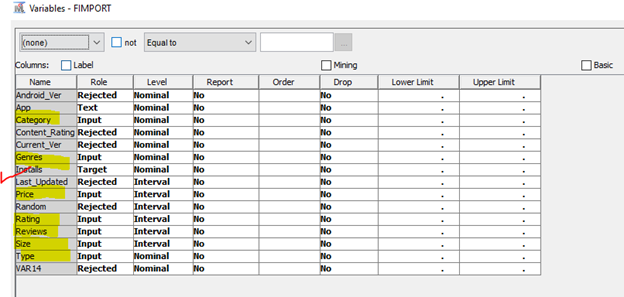
**Models Applied**

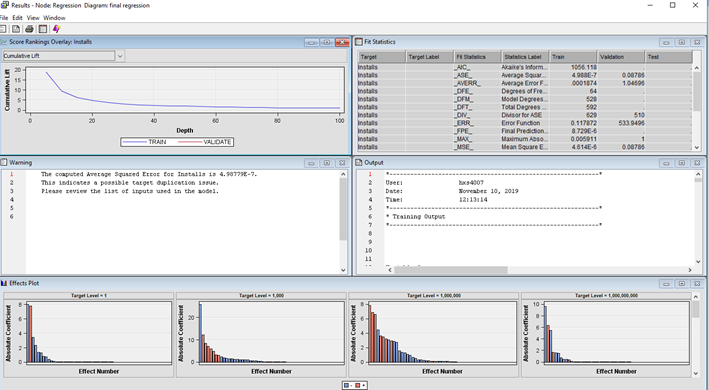
**Linear Regression**

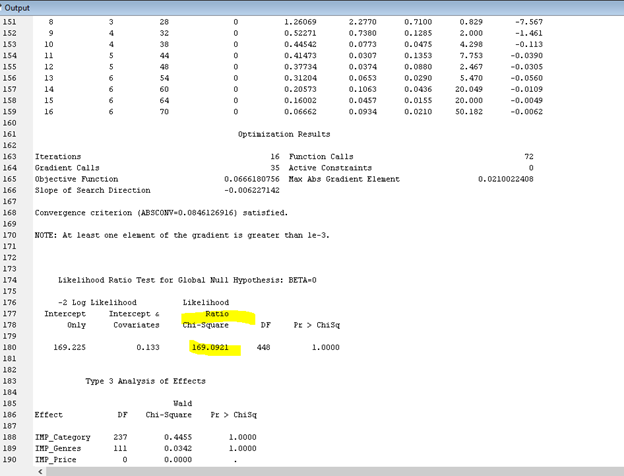
When you run regression on SAS Enterprise Miner, you will get the results like the following observations. The Output table tells you the coefficients of each predictor variable. If the coefficients are positive, it means that that predictor variable has a positive relationship with the target variable.



Variables Considered:

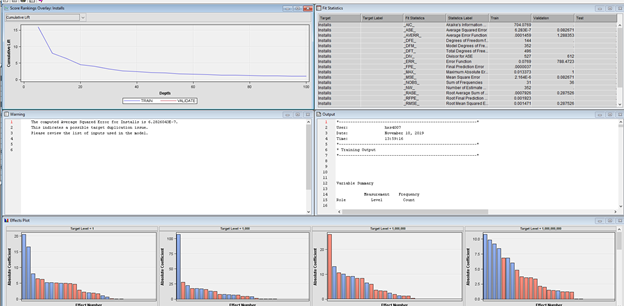






The Chi square value is high, so the maximum likelihood of estimation is good.

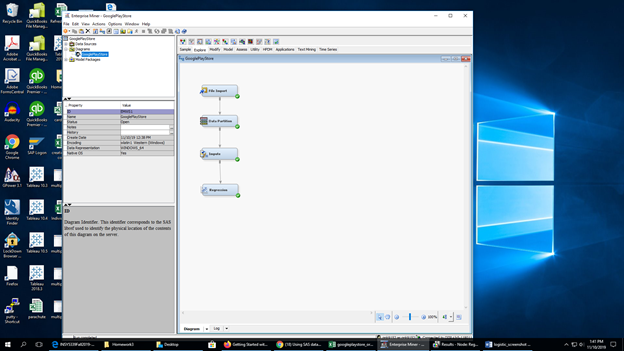
Now considering only the variables which are in our control i.e variables that can be manipulated we ran the regression model again

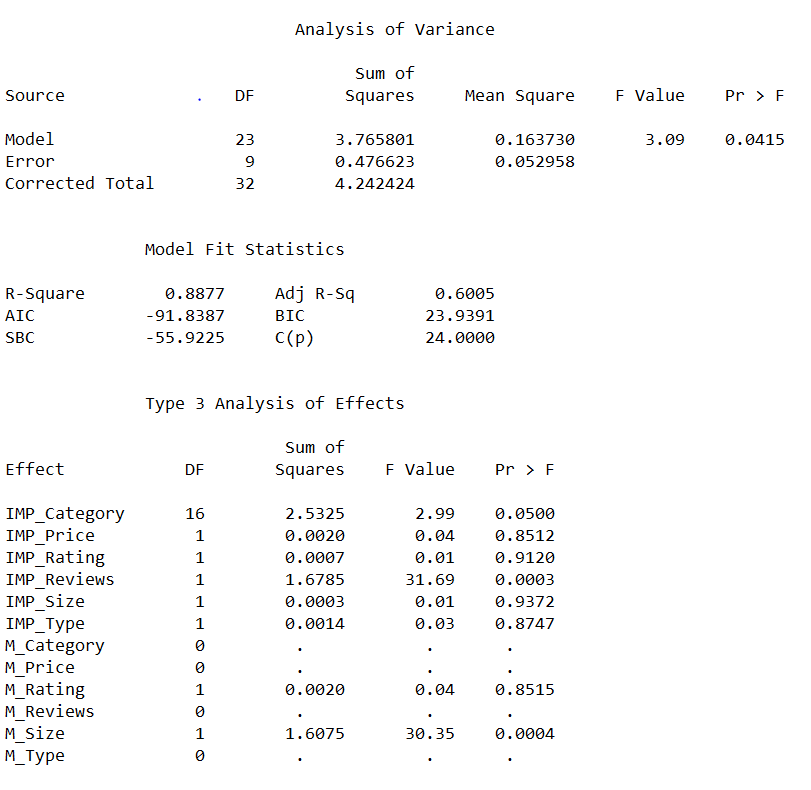


**Logistic Regression**

We ran logistic regression model by first converting number of installs into Binary format. We kept 10M installs as the threshold and the values that are below 10M were converted to Binary 0 and value above 10M were converter to Binary 1.

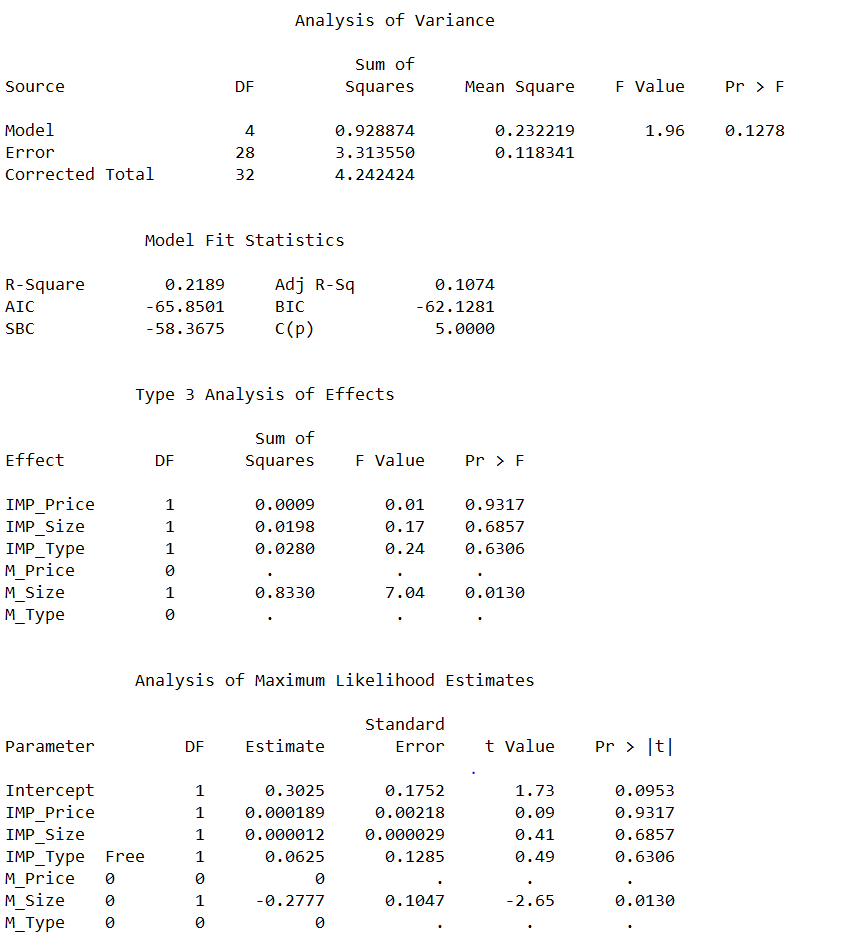
We received the following output.





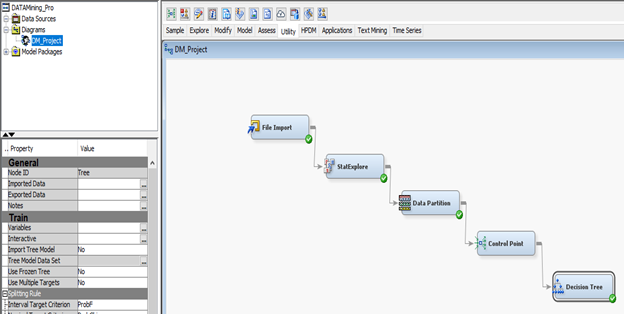
Since the p value is less than the f value we can say that independent variables can be used to predict the dependent variable.

**Using Variables that are in our control.**

****

Here too the p value is less than the f value hence we can conduct the prediction using the independent variables.

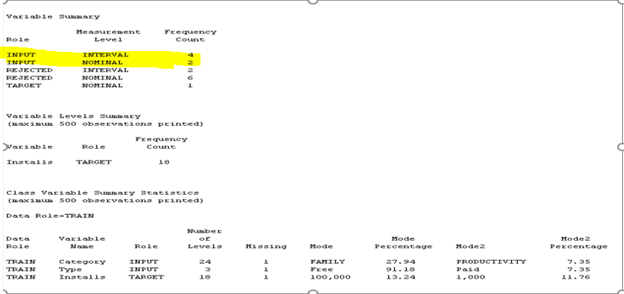
Decision Tree

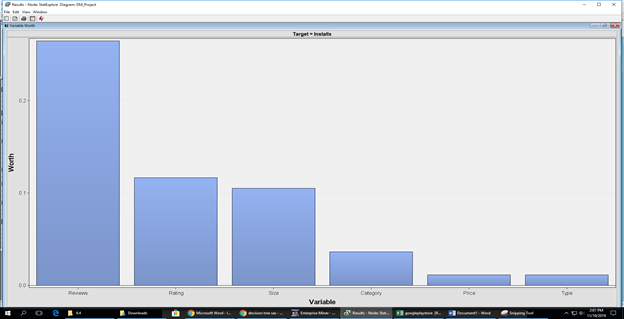


The above Diagram shows the decision tree approach of the project.

Data has been partitioned as training and validation in the ratio 6:4.Leaf size for the decision tree is set as 8.

**Variable Summary :**

****



From the above graph of variable worth it is clear that, Review is contributing more for the analysis of number of installations

