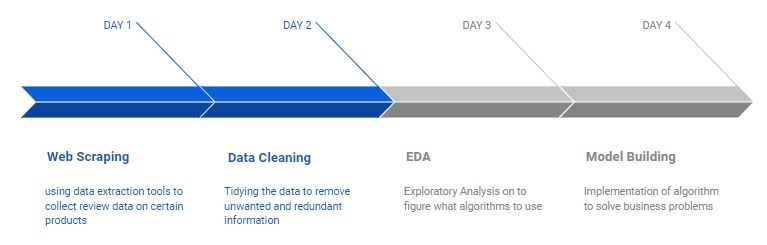
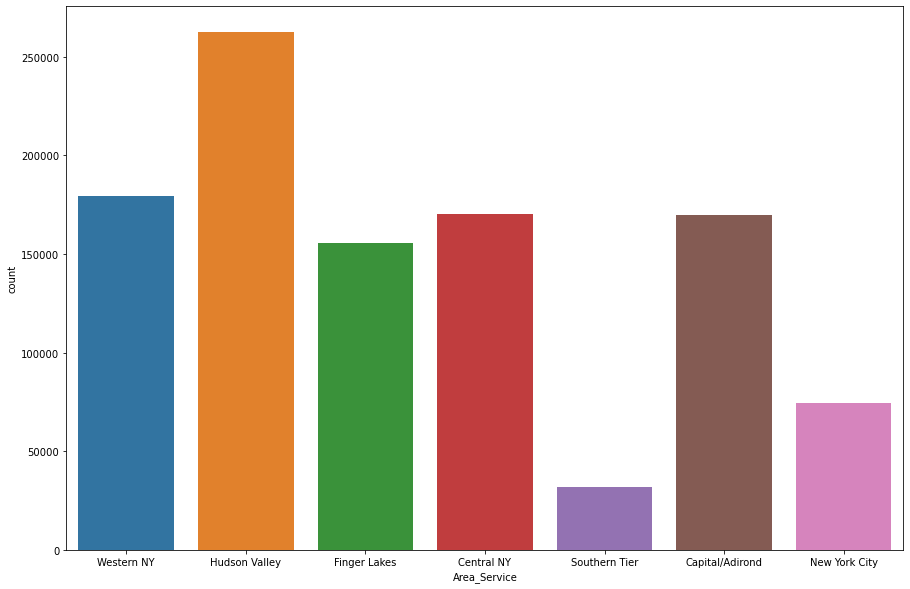
FRAUD ANALYTICS

**By:-NAVEEN**

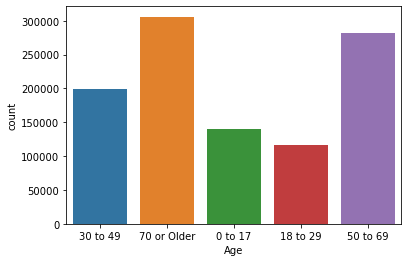
**BUSINESS** **OBJECTIVE**

To Predict whether the customer claim is Genuine or Fraud

PROJECT ARCHITECTURE/FLOW

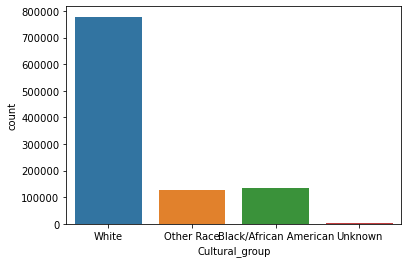
This is a count plot for the area\_service column. We can see that Hudson Valley is the most served area followed by Western NY, Central NY, Capital and Finger Lakes respectively where the data on hospital facilities was collected.

From this plot, we observe highest peak (count) in Hudson Valley and the lowest in the Southern tier.

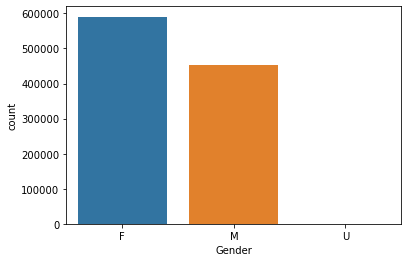
Here is a count plot for the age column which shows patients of different age groups starting from 0 to 70.

We can see that most of the patients are of age group 70 or older followed by group 50 to 69. The minimum amount of patients are in group 0 to 17, followed by group 18 to 29.

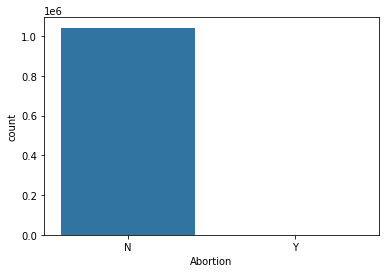
The highest peak is observed in ages groups 70 or older and lowest in the age groups between 18-29.

This is a count plot for the cultural\_group column. Black depicts African Americans, White depicts Americans and the Other Race is the group other than these two.

We can see that the most number of patients are White. Blacks and Other Races are almost equal in number.Unknown are almost negligible.

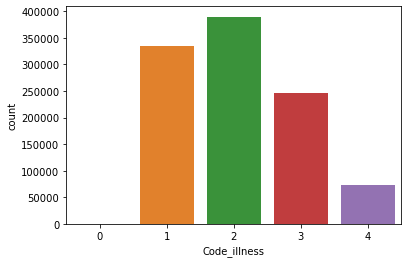
Here is a count plot for the Gender column which shows the gender of the patients.

We can see that female patients are more in number than the male patients.

Here is a count plot for the abortion column.

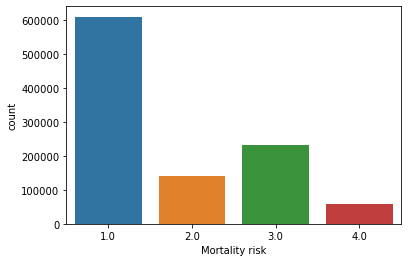
We can see that none of the patients went for abortion.

This depicts that there is an imbalance in the data which could create a bias in the model building.

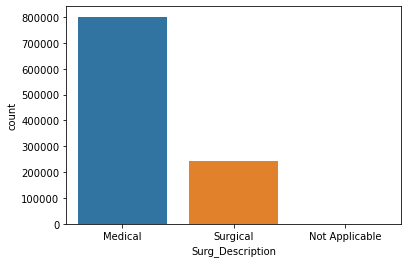
Here is a count plot for the illness code column. 1 depicts mild illness, 2 is for moderate, 3 for severe and 4 for indeterminate.

We can see that most number of patients have moderate illness followed by mild illness & severe.

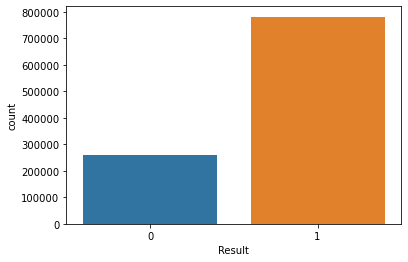
The least number lies in indeterminate illness group.

Here is a count plot for the mortality risk column which shows the mortality risk for the patients. 1 is for minor risk, 2 for moderate, 3 for major and 4 for severe.

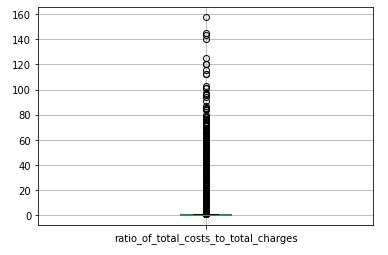
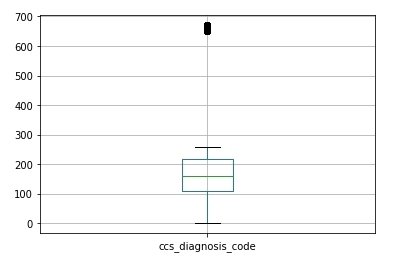
We can see that risk is minor for most of the patients followed by major risk.

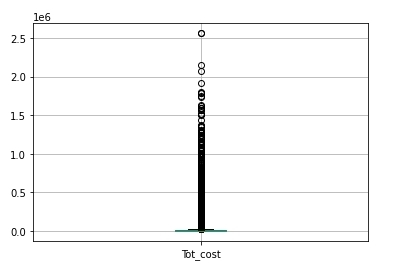
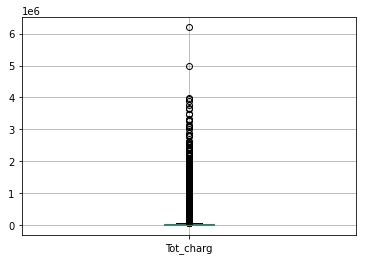
This is the count plot for the Surg\_Description column which shows the different types of treatment provided to patients.

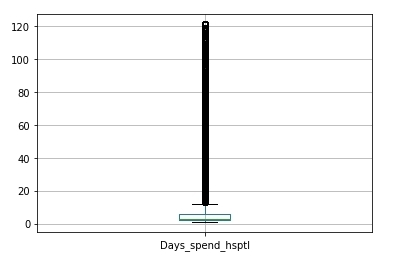
We can see that more amount of patients were given medical treatment as compared to surgical treatment.

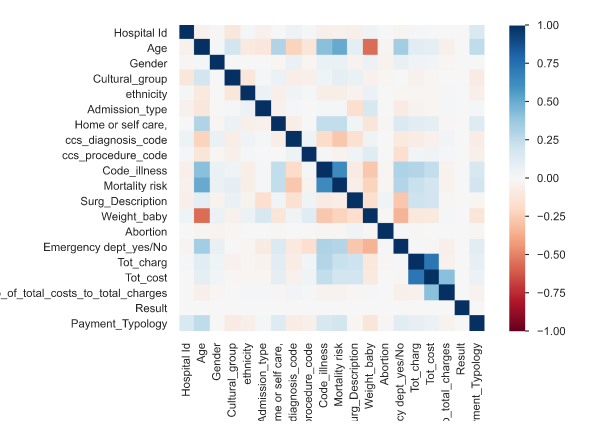
Here is a count plot for the result column. Here 1 shows the count of genuine claims and 0 shows the count of fraud claims.

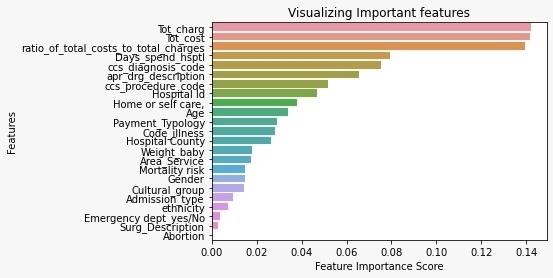
We can clearly see the imbalance in the result as the count of genuine claims is much greater than the count of fraud claims.







Heatmap of Correlation

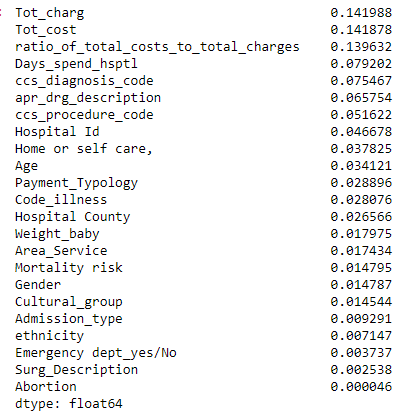
**FEATURE** **ENGINEERING**

• We remove the columns Hospital County and Hospital ID since irrespective of which hospital it is, there will be claims.

• We remove the weight\_baby column to as 90% of the rows are zeros and also it has low score as predicted by the ExtraTrees model.

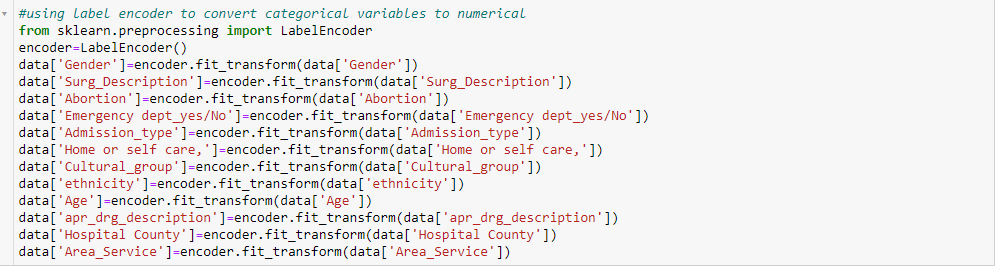
• We remove the apr\_drug\_description as it is just the description of the disease and does not have a great impact in the model building.

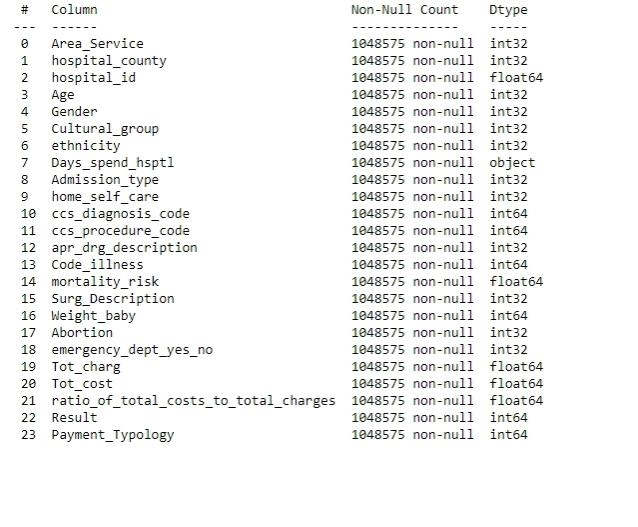
• We remove the ratio\_of\_total\_cost\_to\_charge column as it just indicates whether the hospital is making profit or not and it would not be an important feature for model building.

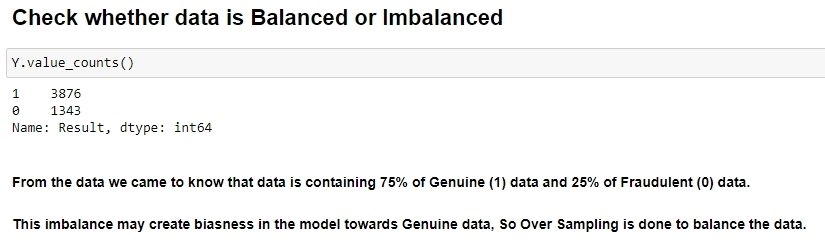
• The feature Abortion has only one value and hence would create biasness in the model, also it has the lowest score as predicted by the ExtraTrees model. Thus we would also drop that column.

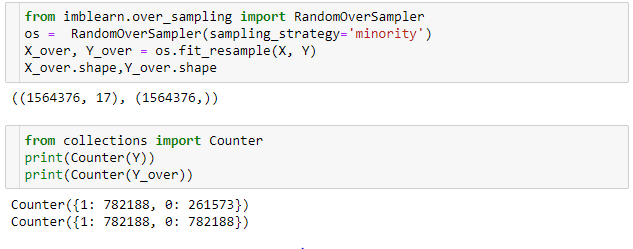
• The observation shows that the Days\_spent\_hsptl column is numerical but it is shown as object type so we use type conversion to convert it into integer type.



• We use Label Encoder to convert the categorical variables into numeric.





Random Over Sampling is done to create the balance in the data, and to remove the bias.

**EDA** **and** **Feature** **Engineering** **Summary**

i. The data set were checked for null values and imputation was used to eliminate them correspondingly.

ii. Different visualizations were plotted to get a deep insight into the data.

iii. Label Encoder was used on the categorical columns to convert them into numeric.

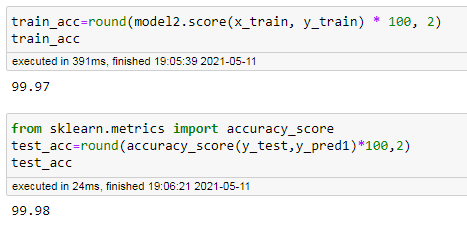
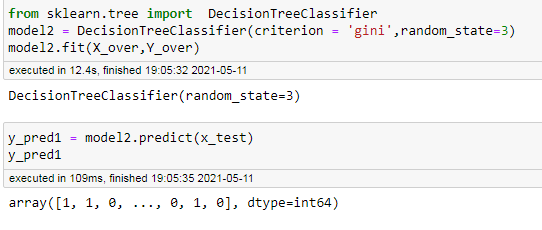
iv. Extra trees classifier was used to get the feature scores and the features which had little impact on the model building were eliminated.

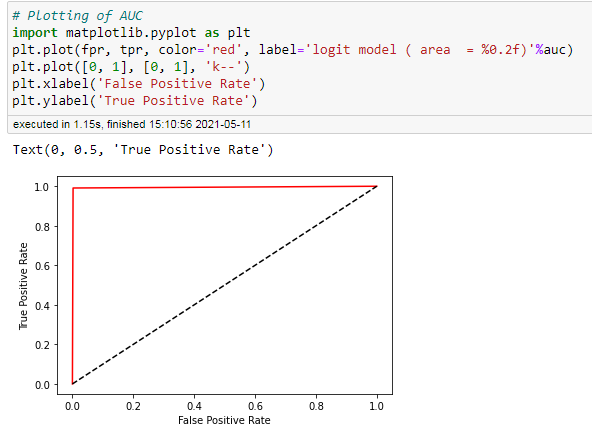
v. Random over sampling was used to create the balance in the result column.

MODEL BUILDING

Comparing accuracies of different models

|  |  |  |
| --- | --- | --- |
| **Model** **Name** | **Training** **Accuracy** | **Testing** **Accuracy** |
| Decision Tree with Entropy Criteria | 99.06% | 99.08% |
| Decision Tree with Gini Criteria | 99.06% | 99.06 |
| Logistic Regression | 74.99% | 74.99% |
| Ada Boost | 50% | 55% |
| Random Forest with under sampling | 50.75% | 48.33% |
| Random Forest with over sampling | 90.5% | 70.83% |
|  |  |  |

Building the Decision Tree Classifier with over sampling as it gave the highest accuracy

ROC Score