**Comparison on the machine learning results**

As it takes more than 10 hours to generate the machine learning HTML file, hence to shorten the run time, the HTML results is prepared based on the subset of the data although the results used in the report & slides are based on full dataset.

|  |  |  |
| --- | --- | --- |
|  | **Report & slides** | **HTML** |
| Data | Full dataset | 30% of full dataset |

Sample checking on the various model performance between models trained with full dataset & models trained with smaller subset of dataset is done to ensure the model results are consistent.

Below are the evidence of checking:

|  |  |  |
| --- | --- | --- |
|  | **Full data** | **30% of data** |
| Comparison on the different models |  |  |
| Comparison on model with data cleaning & without data cleaning |  |  |
| Comparison on directly model on actual claim cost and claim difference between initial and actual |  |  |
| Comparison on full model and simpler model |  |  |
| Comparison on different text mining method |  |  |