Business Problem:

Business analyst and Project Manager discussed with Client regarding the Business Problem.

“To predict whether the claim filed by the Hospital to the Insurance Company is Fraudulent or Genuine.” In Malaysia, government gives health benefits to the Locals (Bhumiputra). Other than Bhumiputra’s, will be insured with other insurance companies. Few of the bhumiputra/insured people tie up with the hospitals and misusing the funds.

Data Analyst and Data Scientist along with BA and PM discussed with SMEs from client side. And the outcome of the discussion was from where we need to get the data to solve the above business problem and what could be the acceptable accuracy percentage (85%) for the final model.

Data Collection:

We get the data from Data warehouse system. Initially built Malaysian Healthcare Data Warehouse system from scratch that is from Requirement Gathering, POC, MyHRDM, Dimensional Modeling, Physical Model, ETL, FFR and Dashboards. Source for the Data warehousing is JSON Formatted files. These files stored in Hadoop environment, from here extract data by using ETL, transform it then load into Data warehouse.

Extracting the data from Data warehousing tables of patient\_dim, visit\_dim, visit\_transaction\_dim, visit\_diagnosis\_dim, visit\_procedure\_dim, payment\_dim, delivery\_dim, delivery\_baby\_dim. Historical data received from csv files. If the claim amount more than cutoff, audit team will do the audit of their claims and will conclude whether the claim filed by the hospital is genuine or fraudulent. Audit team send this information in csv formatted files.

Historical dataset has 28 columns and 10,48,575 rows. Following are the list of columns:

1. Area\_Service
2. HospitalCounty
3. HospitalId
4. Age
5. Gender
6. Cultural\_group
7. Ethnicity
8. Visit\_id
9. Days\_spend\_hsptl
10. Admission\_type
11. Homeorselfcare
12. ccs\_diagnosis\_code
13. ccs\_procedure\_code
14. disease\_description
15. Code\_illness
16. Mortalityrisk
17. Surg\_Description
18. Weight\_baby
19. Abortion
20. Emergencydept\_yes/No
21. Tot\_charg
22. Tot\_cost
23. ratio\_of\_total\_costs\_to\_total\_charges
24. Result
25. Payment\_Typology
26. Certificate\_num
27. zip\_code\_3\_digits
28. year\_discharge

Day-to-day data extracts from Data warehouse system

Audit data extracts from csv files.

Exploratory Data Analysis:

In the EDA, join the patient information and audit data.

Following are the steps performed to impute missing values.

1. For Days\_spend\_hsptl column, replaced 120+ value with 120.
2. For Code\_illness, Tot\_charg, Tot\_cost columns, applied KNN algorithm to impute missing values.
3. For ethnicity column, impute missing values with MODE.
4. In a single datapoint, if 5 features or more are null values, removed those datapoint.
5. For age column, applied Multiple Linear Regression.
6. Converted Object datatypes into float64 or int64
7. Removed the duplicate datapoints.
8. Rename the column names.
9. For numerical columns of Tot\_charg, Tot\_cost, checked the whether the data fall under Normal Distribution or not? If the data is not normally distributed, converted them into Logarithmic Transformation, Reciprocal Transformation, Square root Transformation, Exponential Transformation and Boxcox Transformation then compare which one has normally distributed. Replaced that Normally distributed values with original column values.
10. Implemented label encoder on dependent feature and one-hot encoder(get\_dummies) for independent categorical features

Feature Engineering: Need to check which set of observations (independent features) are more correlated with Y variable.

Following are the steps:

1. For numerical features
   1. Applied the Pearson Correlation by using heatmap function, to find importance of independent features with dependent feature and checked multi-collinearity
   2. Applied SelectKBest method with score\_func function as “f\_regression” or “mutual\_info\_regression”.
2. For categorical features
   1. Applied SelectKBest method with score\_func function as “chi2” or “f\_classif” or “mutual\_info\_classif”.
3. Recursive Feature Elimination
4. With the pps score function we found the correlation within X features (categorical and numerical). For those pps score value greater than 70, excluded those features.
5. ExtraTreesClassifer, ~~SelectFromModel~~ can be used for all the features

When comparing to RFE, SelectFromModel is less robust.

1. Recursive Feature Elimination (rfe) used with algorithm name and hyperparameter values for all the features
2. Visualization techniques
   1. Boxplot
   2. Barplot
   3. Crosstab

After applying above techniques finalized the important independent features which are more influencing dependent feature

Model Building:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S. No. | Algorithm Name (CV =10) | Accuracy without SMOTE | Accuracy with SMOTE | Precision Accuracy | Recall Accuracy | KFold (cv=10) |
| 1 | Logistic Regression | 75% | 64% | 99% | 75% | 75% |
| 2 | RandomForestClassifier | 67% | 55% | 83% | 75% | 67% |
| 3 | DecisionTreeClassifier | 62% | 55% | 75% | 75% | 62% |
| 4 | XGBClassifier | 75% | 59% | 99% | 75% | 75% |
| 5 | GradientBoostingClassifier | 75% | 65% | 99% | 75% | 75% |
| 6 | KNN | 70% | 55% | 90% | 75% | 75% |
| 7 | SVM | 75% | 58% | 100% | 75% | 75% |
| 8 | NaiveBayes (GaussianNB) | 75% | 41% | 99% | 75% | 75% |
| 9 | NaiveBayes (MultinomialNB) | 74% | 53% | 99% | 75% | 72% |

Imbalance Data Techniques:

Model Validation:

Before model creation we will apply cross validation

Leave one out cross validation(LooCV): if we have 1000 records, it consider first data point as test and remaining will be in training, based on this model will train on this particular data and tested on one (first data point) single data. Similarly, in the next experiment the next data point (second data point) will be your test and remaining will be training data again same process will be continued 1000 times. If we want to cover every data point, must run 1000 times.

Disadvantages:

* Run many iterations
* Lead to low bias (For the training dataset you get better result, when you get new data in the production, then the accuracy goes down)

K-Fold Cross Validation:

In this if your dataset is 1000, if we select k value that is 5, means it has 5 experiments. For each experiment it will decide based on k value how many data points will be there (1000/5 = 200) in the test data. In the first experiment first 200 will be the test data and the remaining will be the training data. The first experiment model will train on 800 dataset and will be tested on 200 datasets then will get our accuracy 1. In the second experiment the next 200 records will be your test data and remaining will be training data then model will be trained with 800 training dataset and will be tested on 200 datasets then will get our accuracy 2. Similarly, will do for the remaining 3 experiments then mean of 5 experiments accuracy then that could be the accuracy of the model.

Disadvantages:

* In my test data or training data, there is the possibility of one class (as in binary classification) of the data. If the data is imbalanced, the model may not give accuracy with the new data.

Stratified Cross Validation:

Everything will be similar like K-Fold cross validation but whenever we select the test data or training data, it is making sure that the number of records will be distribution in good proportion of the whole dataset so that your model will give good accuracy.

Deployment: