ITSM - ML Project

PROJECT SUMMARY:

The project goal is to create a predictive model which predicts High priority and Low priority tickets, so that they can take preventive measures or fix the problem before it surfaces.

The dataset has the 1200 entries which is used to perform machine learning. This machine learning problem is Supervised Classification problem. There are 25 features in the dataset. The Shape of the data is 46606x25. These features are classified into Quantitative (Numerical) and Qualitative (Categorical) data where 13 features are qualitative, 6 features are quantitative and 6 features were alphanumerical data.

Data Processing is done to check for missing values and null values in the dataset. Data exploration is done to check the distribution analysis between the feature variables. The machine learning model which is used in this project is Random Forest, XG Boost, Decision Tree, SVM, K-Nearest Neighbor, MLP Classifier (ANN) and Logistic Regression where XG Boost classifier predicted higher accuracy of 87% with Recall rate of 84%.

Label Encoding is done to convert the categorical data into numerical data, because, the machine learning methods are based on numerical values. The overall project was performed using machine learning model, Sampling methods and visualization techniques.

Data Analysis:

Data Analysis is done by describing the features in the data. There are 25 features in the dataset. The Shape of the data is 46606x25. These features are classified into Quantitative (Numerical) and Qualitative (Categorical) data where 13 features are qualitative, 6 features are quantitative and 6 features were alphanumerical data.

Categorical Features:

In categorical features, there are the nominal, ordinal, ratio, or interval based values. The categorical features are as follows,

- CI_Cat
- CI_Subcat
- Status
- Impact
- Urgency
- Priority
- Category
- Alert_Status
- Closure_Code
- No_of_Reassignments
- No_of_Related_Interactions
- No_of_Related_Incidents
- No_of_Related_Changes

Numerical Features:

In numerical features, there are the discrete, continuous, or timeseries based values. The numerical features are as follows:

- number_cnt
- Open_Time
- Close_Time
- Reopen_Time
- Resolved_Time
- Handle_Time_hrs

Alphanumeric Features: This feature has numerical and alphanumeric data.

- CI_Name
- WBS
- Incident_ID
- KB_number
- Related_Interaction
- Related_Change

Data Cleaning:

The Data cleaning/ Data wrangling is the main part of the Data science project because the missing data/ null values in the dataset will lead to the lower accuracy. If the data is already structured and cleaned, there is no need for data cleaning. In the given dataset, there were significant wrong entries and missing values which must be handled.

- **Existing dataset:**
  The following data is the original dataset, having **46606** rows (records) with some missing values and **25** columns (feature variables).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46606 entries, 0 to 46605
Data columns (total 25 columns):
CI_Name                       46606 non-null object
CI_Cat                        46606 non-null object
CI_Subcat                     46606 non-null object
WBS                           46606 non-null object
Incident_ID                   46606 non-null object
Status                        46606 non-null object
Impact                        46606 non-null object
Urgency                       46606 non-null object
Priority                      46606 non-null object
number_cnt                    46606 non-null object
Category                      46606 non-null object
KB_number                     46606 non-null object
Alert_Status                  46606 non-null object
No_of_Reassignments           46606 non-null object
Open_Time                     46606 non-null object
Reopen_Time                   46606 non-null object
Resolved_Time                 46606 non-null object
Close_Time                    46606 non-null object
Handle_Time_hrs               46606 non-null object
Closure_Code                  46606 non-null object
No_of_Related_Interactions    46606 non-null object
Related_Interaction           46606 non-null object
No_of_Related_Incidents       46606 non-null object
No_of_Related_Changes         46606 non-null object
Related_Change                46606 non-null object
dtypes: object(25)
memory usage: 8.9+ MB
```

❖ Refined dataset:

- As some of the columns are not useful for Analysis and Modelling, it's better to drop them.
  Dropped columns are - 'Status', 'number_cnt', 'Alert_Status', 'Open_Time', 'Resolved_Time', 'Close_Time', 'Handle_Time_hrs', 'Related_Interaction', 'Related_Change'.
- As per my analysis, I felt there were Nan values in CI_Cat and CI_Subcat for corresponding CI_name (OVR). There were Nan values for CI_Name 'OVR', So I dropped the rows which are having null values in the CI_Cat and CI_Subcat.
- As Priority is based on Impact and Urgency, we cannot fill the missing/NaN values for these 3 features. (Impact,Urgency,Priority)
- As **'No_of_Related_Changes**' and **'No_of_Related_Incidents'** columns has more number of missing values and also not useful for prediction, we can drop the column.
- As there are more Null values in CI_name as 'CBD', we will have to analyze the Closure_code values for corresponding CI_Name 'CBD' and fill the values. After analyzing, the Closure_code in 'Hardware' have more number of counts among the CI='CBD', so we can replace the Nan values in the Closure_code to 'Hardware'.
- As we are focusing only on the 'incident' category, we can remove other categories.
- Adding a new column based on **'Reopen_Time'** as **'ReOpen_flag'** for the incidents which has been Re-Opened ➔ (1) or Not Re-Opened ➔ (0).
- As per the business problem, we need to predict High priority tickets i.e., Priority 1 and Priority 2.
  So we filter the feature variable **'Priority'** based on condition and add a new feature as **'New_Priority'** ➔ Target variable (y)
  'New_Priority' will contain 'High_Priority' and 'Low_Priority'.
  The Priority which is 1 and 2 are High_Priority.
  The Priority which is 3, 4 and 5 are Low_Priority.

Data Exploration (EDA):

Data exploration is done by visualizing the dataset to get the proper insights. I have performed data analysis by the visualizing the distribution of some of the features.

- Distribution Analysis: Here I checked the distribution of the data, so that we can analyze how the features are distributed between other features. The distribution analysis is done for both numerical and categorical features. This will give some insights about where the majority of the data is distributed.

    - The distribution of Application category in the dataset is more when compared to other categories.
    - Among the distribution of sub-categories, Web based application and Server based application is more when compared to other sub-categories.
    - Most of the distribution of number of related interactions and re-assignments occurred from 0 to 2.
    - 96% of the incidents have not Re-opened whereas only 4% were Re-opened incidents.
    - The distribution of priority is more in Priority- 4 (i.e., 61%), and remaining priority i.e., 5(23%), 3(14%), 2(2%), 1(0%).

Machine Learning Model implementation:

As the data here is more skewed towards low priority tickets and number of data entries are more among low priority than high priority tickets. There is an imbalance nature in the dataset. So here we used under-sampling technique to equalize the dataset among high priority and low priority tickets.

Before Under-sampling:

New priority	Total count
Low Priority	35634
High Priority	685

After Under-sampling:

New priority	Total count
Low Priority	685
High Priority	685

The machine learning algorithms used are best for classification and labelled data. The train and test data are divided and trained in the model and passed through the machine learning algorithm for prediction.

The Accuracy, Recall and Precision given by different are as follows:

Algorithm	Accuracy	Recall	Precision
Logistic Regression	76	73	76
K-Nearest Neighbor	83	79	86
SVM	83	80	84
MLP Classifier (ANN)	79	60	93
Decision Tree	83	67	96
Random Forest	86	79	91
XG Boost	87	84	89

Result:

Algorithm XG boost classifier has given a good Accuracy with 87% and Recall rate with 84%, which is more efficient for Model prediction.