MID TERM ASSIGNMENT

NAME: POOJAN VADALIYA (1281587)

COURSE CODE : DATA*6300: ANALYSIS OF BIG DATA

MID-TERM HOME ASSIGNMENT

DATASET :

Diabetes 130-US hospitals for years 1999-2008
Sample Count : 101766 samples
Feature Count : 47 Features

Problem Statement :

The problem perspective considered here is to find out how many patients were readmitted in the hospital.

DATA PREPROCESSING :

- After Loading the dataset features with **missing values** were identified.
- Features with missing values greater than 40% were dropped which included 'weight', 'medical_specialty' , 'payer_code'.
- Type of other features with missing values were checked followed by the Ordinality for the categorical features, where observation was made for the following.
    - **race :** It was an non ordinal categorical value with 5 categories which was handled by considering all the null values as 'Other' category.
    - **Diag_1, Diag_2, Diag_3 :** It also consisted of non-ordinal categorical values.
- **Handling Diagnostics Feature:** Considering excessive frequency of categories for all three diagnostics columns, all the categories were re-grouped into 10 categories by referring the ICD-9 code table.
    - All the categories were mapped into 10 group names according to the icd9 codes described in the table.
    - Missing values were then imputed by the model of the categories of each feature.
    - And considering the non-ordinal nature of each column all the three categories were one hot encoded to convert the categorical values into a numerical value for data modelling.

Group name	icd9 codes	Number of encounters	% of encounter	Description
Circulatory	390–459, 785	21,411	30.6%	Diseases of the circulatory system
Respiratory	460–519, 786	9,490	13.6%	Diseases of the respiratory system
Digestive	520–579, 787	6,485	9.3%	Diseases of the digestive system
Diabetes	250.xx	5,747	8.2%	Diabetes mellitus
Injury	800–999	4,697	6.7%	Injury and poisoning
Musculoskeletal	710–739	4,076	5.8%	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580–629, 788	3,435	4.9%	Diseases of the genitourinary system
Neoplasms	140–239	2,536	3.6%	Neoplasms
	780, 781, 784, 790–799	2,136	3.1%	Other symptoms, signs, and ill-defined conditions
	240–279, without 250	1,851	2.6%	Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
	680–709, 782	1,846	2.6%	Diseases of the skin and subcutaneous tissue
	001–139	1,683	2.4%	Infectious and parasitic diseases
	290–319	1,544	2.2%	Mental disorders
Other (17.3%)	E–V	918	1.3%	External causes of injury and supplemental classification
	280–289	652	0.9%	Diseases of the blood and blood-forming organs
	320–359	634	0.9%	Diseases of the nervous system
	630–679	586	0.8%	Complications of pregnancy, childbirth, and the puerperium
	360–389	216	0.3%	Diseases of the sense organs
	740–759	41	0.1%	Congenital anomalies

- **Handling Other Categorical Features:** All the categorical features were identified and categories ordinality were checked.
- Categorical Features were handled in the following manner:
    - **Medicinal drugs :** Here each drug category was label encoded considering 'Down' : -1, 'No' : 0, 'Steady' : 1,' Up' : 2
    - **Age**: Each age category was labeled into the average value of each age category.
    - **Race :** As being non ordinal it was one hot encoded to convert the categories to numerical value.

- **Max_glu_serum:** Here each category was label encoded consider None : 0, Norm : 1 , >200 and >300 : '1'.
- **A1Cresult :** Here each category was label encoded consider None : 0, Norm : 1 , >7 and >8 : '1'.
- **readmitted:** After checking the target column three categories were grouped into 0 and 1 were 'NO' was considered as 0 and '<30' and '>30' as 1 which signified whether whether a patient was readmitted again or not in the hospital.

CLASSIFICATION MODELS CONSIDERATION AND RESULTS

MODEL NAME		PRECISION	RECALL	F1-SCORE	ACCURACY
Naïve Bayes	0	0.63	0.73	0.68	0.62
	1	0.61	0.50	0.55	
Logistic Regression	0	0.62	0.80	0.70	0.62
	1	0.63	0.41	0.50	
Decision Tree	0	0.60	0.58	0.59	0.56
	1	0.52	0.54	0.53	
Support Vector Classifier	0	0.59	0.90	0.71	0.61
	1	0.70	0.27	0.39	
Random Forest	0	0.65	0.74	0.69	0.64
	1	0.63	0.53	0.58	
Bagging Classifier	0	0.62	0.72	0.67	0.60
	1	0.59	0.47	0.52	
Neural Network	0	0.54	1.00	0.70	0.54
	1	0.00	0.00	0.00	
XG Boost Classifier	0	0.66	0.72	0.69	0.65
	1	0.63	0.57	0.60	
Cat Boost	0	0.64	0.76	0.70	0.64
	1	0.64	0.50	0.56	

CONCLUSION:

After observing the above accuracy tables XG Boost Classifier can be consider as the best model of all followed by random forest and CatBoost.

Consideringflexibility and ability to capture complex relationships in the data. XGBoost helped to handle both linear and non-linear relationships effectively, allowing it to model intricate patterns in the dataset which resulted to into better accuracy and classification of the dataset.