Data Challenge National Data Management Center

Data Analytics, Modeling, and Visualization Division

By Naol Lamesa
06 Jun, 2024

Business Objective

The objective of this data challenge is to assess the interns' analytical and problem-solving skills by presenting them with real-world problems that organizations face daily. This challenge aims to evaluate the interns' proficiency in data science modeling, including their mathematical, programming, and conceptualization abilities.

The Child Health and Mortality Prevention Surveillance (CHAMPS) dataset has been provided for this challenge. The purpose of the CHAMPS dataset is to collect, analyze, and share data to help identify the causes of child deaths in areas with high child mortality. The dataset includes information on underlying causes of death from the infant's side and maternal factors contributing to these deaths. By engaging with this dataset, interns will gain valuable experience in addressing complex data challenges that have significant real-world implications.

Key Findings

Exploratory Data Analysis

Based on the given dataset and the decoded variables, I performed the following preprocessing and Exploratory Data Analysis (EDA).

```
# a. read dataset
df=load_dataset.load_data(path)
# b. number of rows and columns in dataset
df.shape
# c. enumerate the columns of dataset
for index, column in enumerate(df.columns):
    print(f"{index}: {column}")

# d. rename the columns
df.rename(columns={'dp_013':'case_type'},inplace=True)

# e. rename the values
df['case_type'] = df['case_type'].replace({'CH00716': 'Stillbirth',
                                           'CH01404': 'Death in the first 24
hours',
                                           'CH01405':'Early Neonate',
                                           'CH01406':'Late Neonate',
                                           'CH00718':'Infant',
                                           'CH00719':'Child'})
# f. Show the proportion of null values in each column.
null_percentage =(df.apply(pd.isnull).sum()/df.shape[0])*100
for column, percentage in null_percentage.items():
    print(f"{column}: {percentage:.2f}%")
```

The data has 444 rows and 381 columns. The column dp_106 contains 100% null percentage, followed by columns qualifier_22 and dp_092.

Descriptive Data analysis

The magnitude and proportion of each of the infant underlying cause for child death.
Top 5:

dp_108	magnitude
Intrauterine hypoxia	148
Birth asphyxia	33
Undetermined	28
Severe acute malnutrition	24
Craniorachischisis	16

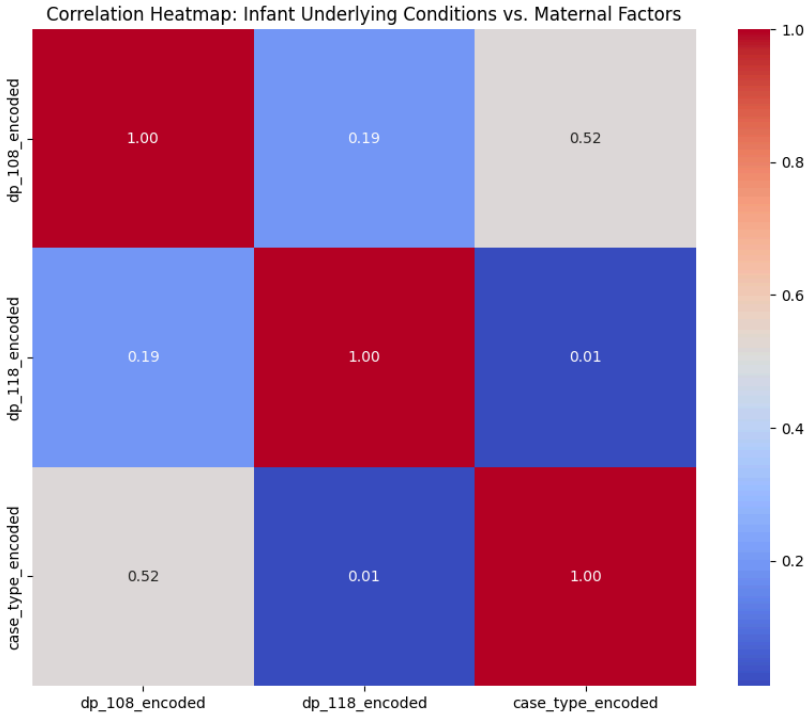
The proportion and magnitude of the maternal factors contributing for child death.
Top 5:

dp_118	magnitude
Preeclampsia	32
Twin pregnancy	12
Abruption placenta	11
Eclampsia	9
Fetus and newborn affected by other forms	5

The proportion of the child death by the case type.
Top 5:

Case_type	magnitude
Stillbirth	53.8
Death in the first 24 hours	15.5
Early Neonate	11
Child	9.4
Infant	6

Correlation analysis



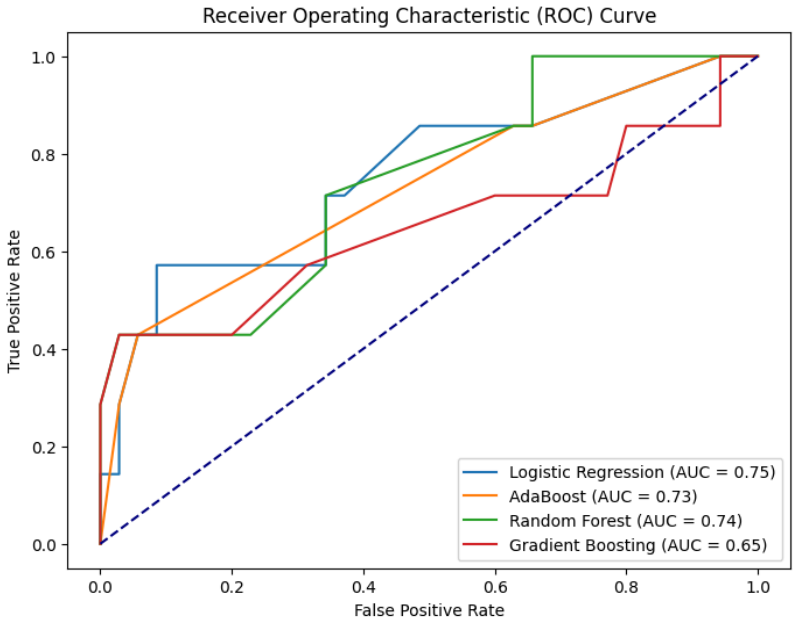
The correlation heatmap illustrates mostly weak or negligible correlations among the features. The standout observation is a positive correlation between 'dp_108_encoded'' and 'case_type_encoded,' indicating that increased infant underlying cause aligns with higher case type.

Feature Engineering

Feature	Logistic Regression	SVM	Adaboost	Random Forest	Gradient Boosting	XGBoost
dp_108_encoded	1.4505	2.00	0.9600	0.6969	0.7661	0.9381
dp_118_encoded	0.0257	0.00	0.0400	0.3031	0.2339	0.0619

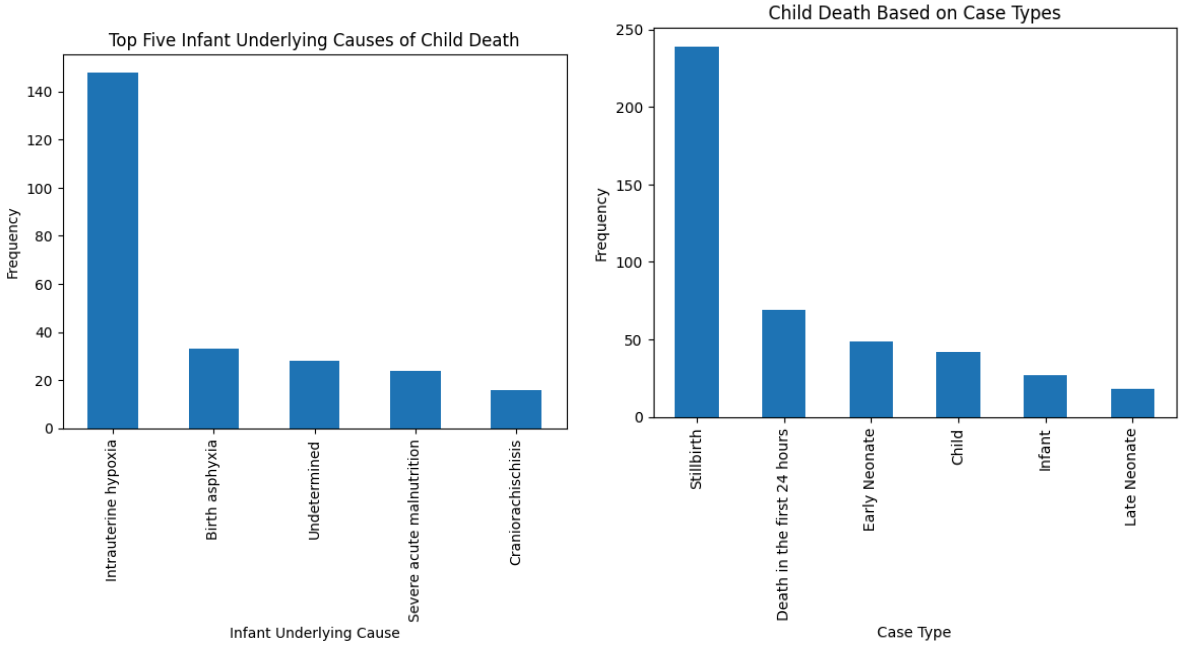
Across all six classification algorithms, the feature `dp_108_encoded` consistently shows higher importance than `dp_118_encoded`. This suggests that `dp_108_encoded` is a more significant predictor for the target variable in this classification task.

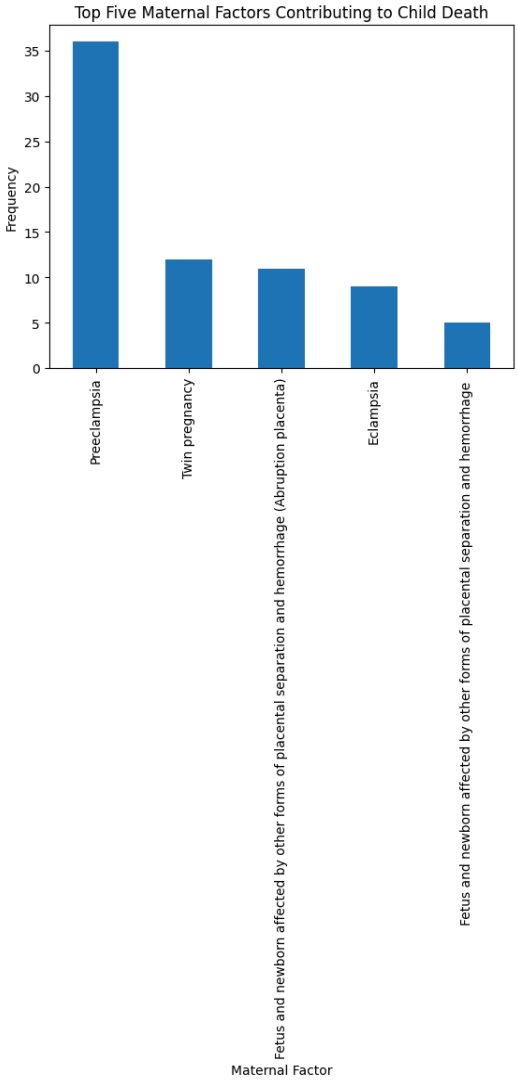
Model Evaluation



Logistic regression shows the highest ROC curve, followed by AdaBoost and Random Forest.

Result Visualization





Conclusion

The feature `dp_108` is identified as a crucial predictor across multiple classification models, indicating its significant role in predicting child death causes. Logistic regression emerged as the most effective model based on ROC curve analysis, highlighting its potential for real-world application in identifying critical factors in child mortality. This challenge has provided valuable insights and practical experience in tackling complex data-driven problems.