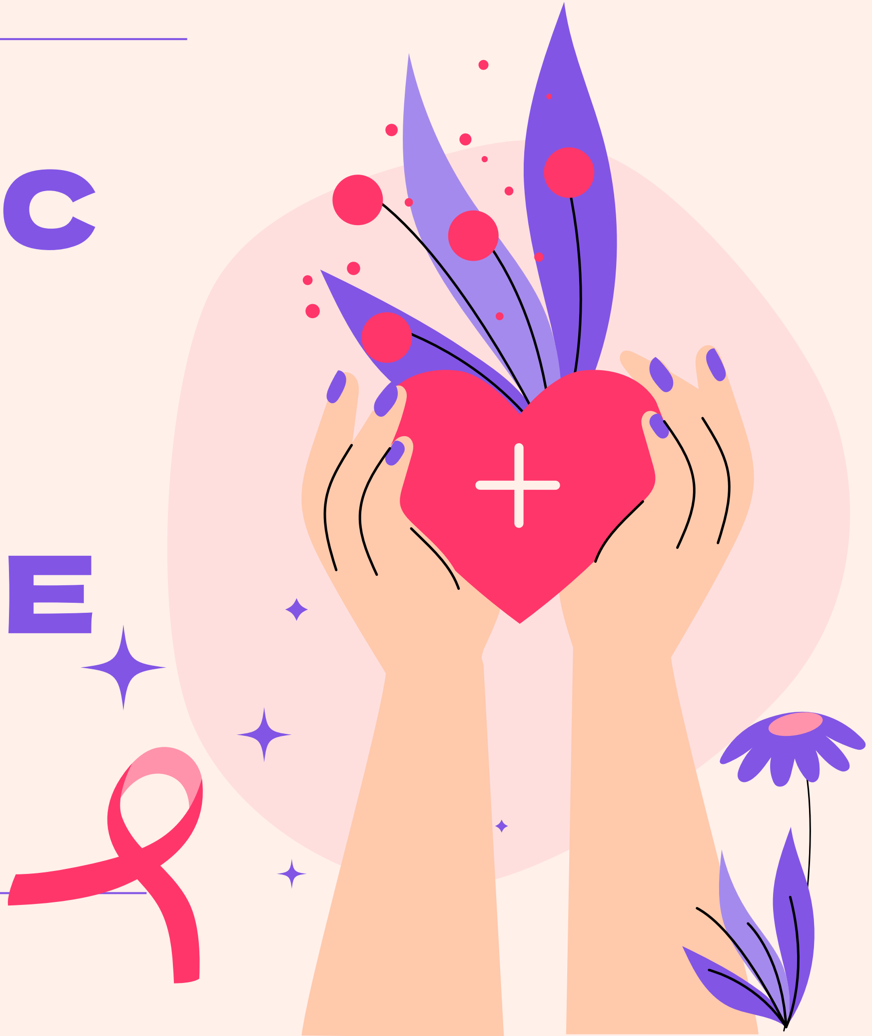
METASTATIC CANCER & EQUITY IN HEALTHCARE

By. Ajibola-Elias Maryam



CONTENTS

Data Introduction

Introduction

- Metastatic TNBC is considered the most aggressive TNBC & requires most urgent and timely treatment. Unnecessary delays in diagnosis & subsequent treatment can have devastating effects in these difficult cancers

- The primary goal is to detect relationships between demographics of the patients and the likelihood of getting timely treatment

- The secondary goal is to see if environmental hazard impact proper diagnosis and treatment.

- However the goal of our model is to just predict whether the patients received metastatic cancer diagnosis with 90 days or not.

Introduction

Columns

- patient_id-unique identification number of patient
- patient_race-Asian, African, American, Hispanic/Latino, White, Other Race
- payer_type-payer type at Medicaid, Commercial, Medicare on the metastatic date
- patient_state- Patient State (e.g. AL, AK, AZ, AR, CA, CO etc…) on the metastatic date
- patient_zip3- Patient Zip3 (e.g. 190) on the metastatic date
- patient_age - Derived from Patient Year of Birth (index year minus year of birth)
- patient_gender - F, M on the metastatic date
- bmi - If Available, will show available BMI information (Earliest BMI recording post metastatic date)
- breast_cancer_diagnosis_code - ICD10 or ICD9 diagnoses code
- breast_cancer_diagnosis_desc - ICD10 or ICD9 code description. This column is raw text and may require NLP/ processing and cleaning
- metastatic_cancer_diagnosis_code - ICD10 diagnoses code
- metastatic_first_novel_treatment - Generic drug name of the first novel treatment (e.g. "Cisplatin") after metastatic diagnosis

Introduction

Columns

- metastatic_first_novel_treatment_type - Description of Treatment (e.g. Antineoplastic) of first novel treatment after metastatic diagnosis
- region - Region of patient location
- division - Division of patient location
- population - An estimate of the zip code's population.
- density - The estimated population per square kilometer.
- age_median - The median age of residents in the zip code.
- male - The percentage of residents who report being male (e.g. 55.1).
- female - The percentage of residents who report being female (e.g. 44.9).
- married - The percentage of residents who report being married (e.g. 44.9).
- family_size - The average size of resident families (e.g. 3.22).
- income_household_median - Median household income in USD.
- income_household_six_figure - Percentage of households that earn at least $100,000 (e.g. 25.3)
- home_ownership - Percentage of households that own (rather than rent) their residence.

Introduction

Columns

- housing_units - The number of housing units (or households) in the zip code.
- home_value - The median value of homes that are owned by residents.
- rent_median - The median rent paid by renters.
- education_college_or_above - The percentage of residents with at least a 4-year degree.
- labor_force_participation - The percentage of residents 16 and older in the labor force.
- unemployment_rate - The percentage of residents unemployed.
- race_white - The percentage of residents who report their race White.
- race_black - The percentage of residents who report their race as Black or African American.
- race_asian - The percentage of residents who report their race as Asian.
- race_native - The percentage of residents who report their race as American Indian and Alaska Native.
- race_pacific - The percentage of residents who report their race as Native Hawaiian and Other Pacific Islander.
- race_other - The percentage of residents who report their race as Some other race.
- race_multiple - The percentage of residents who report their race as Two or more races.

Introduction

Columns

- hispanic - The percentage of residents who report being Hispanic. Note: Hispanic is considered to be an ethnicity and not a race.
- age_under_10 - The percentage of residents aged 0-9.
- age_10_to_19 - The percentage of residents aged 10-19.
- age_20s - The percentage of residents aged 20-29.
- age_30s - The percentage of residents aged 30-39.
- age_40s - The percentage of residents aged 40-49.
- age_50s - The percentage of residents aged 50-59.
- age_60s - The percentage of residents aged 60-69.
- age_70s - The percentage of residents aged 70-79.
- age_over_80 - The percentage of residents aged over 80.
- divorced - The percentage of residents divorced.
- never_married - The percentage of residents never married.
- widowed - The percentage of residents never widowed.
- family_dual_income - The percentage of families with dual income earners.

Introduction

Columns

- income_household_under_5 - The percentage of households with income under $5,000.
- income_household_5_to_10 - The percentage of households with income from $5,000-$10,000.
- income_household_10_to_15 - The percentage of households with income from $10,000-$15,000.
- income_household_15_to_20 - The percentage of households with income from $15,000-$20,000.
- income_household_20_to_25 - The percentage of households with income from $20,000-$25,000.
- income_household_25_to_35 - The percentage of households with income from $25,000-$35,000.
- income_household_35_to_50 - The percentage of households with income from $35,000-$50,000.
- income_household_50_to_75 - The percentage of households with income from $50,000-$75,000.
- income_household_75_to_100 - The percentage of households with income from $75,000-$100,000.
- income_household_100_to_150 - The percentage of households with income from $100,000-$150,000.
- income_household_150_over - The percentage of households with income over $150,000.
- income_individual_median - The median income of individuals in the zip code.

Introduction

Columns

- poverty - The median value of owner occupied homes.
- rent_burden - The median rent as a percentage of the median renter's household income.
- education_less_highschool - The percentage of residents with less than a high school education.
- education_highschool - The percentage of residents with a high school diploma but no more.
- education_some_college - The percentage of residents with some college but no more.
- education_bachelors - The percentage of residents with a bachelor's degree (or equivalent) but no more.
- education_graduate - The percentage of residents with a graduate degree.
- education_stem_degree - The percentage of college graduates with a Bachelor's degree or higher in a Science and Engineering (or related) field.
- self_employed - The percentage of households reporting self-employment income on their 2016 IRS tax return.
- farmer - The percentage of households reporting farm income on their 2016 IRS tax return.
- disabled - The percentage of residents who report a disability.
- limited_english - The percentage of residents who only speak limited English.
- commute_time - The median commute time of resident workers in minutes.

Introduction

Columns

- health_uninsured - The percentage of residents who report not having health insurance.
- veteran - The percentage of residents who are veterans.
- ozone - Annual Ozone (O3) concentration data at Zip3 level. This data shows how air quality data may impact health.
- PM25 - Annual Fine Particulate Matter (PM2.5) concentration data at Zip3 level. This data shows how air quality data may impact health.
- NO2 - Annual Nitrogen Dioxide (NO2) concentration data at Zip3 level. This data shows how air quality data may impact health.

Target

- DiagPeriodL90D - Diagnosis Period Less Than 90 Days. This is an indication of whether the cancer was diagnosed within 90 Days.

Data Understanding

Understanding the data

- Target: 1 - diagnosed within 90 day ; 0-didn't get diagnosed within 90 days
- Binary classification
- Three csv files: Train, Test, and sample_submission
- Size: 16.36MB
- 83 columns, 12906 rows
- 8327 missing values
- No duplicates
- Numerical & categorical values
- No Invalid Entries

Understanding the data

Outliers Detected

- The mean being bigger than the max values, suggested that that we might have outliers in the dataset
- Will be handles with log transformation..

```
df.describe()
```

	patient_id	patient_zip3	patient_age	bmi	population	density
count	12906.000000	12906.000000	12906.000000	3941.000000	12905.000000	12905.000000
mean	547381.196033	573.754300	59.183326	28.984539	20744.441237	1581.950419
std	260404.959974	275.447534	13.335216	5.696906	13886.903756	2966.305306
min	100063.000000	101.000000	18.000000	14.000000	635.545455	0.916667
25%	321517.000000	331.000000	50.000000	24.660000	9463.896552	171.857143
50%	543522.000000	554.000000	59.000000	28.190000	19154.190480	700.337500
75%	772671.750000	846.000000	67.000000	32.920000	30021.278690	1666.515385
max	999896.000000	999.000000	91.000000	85.000000	71374.131580	21172.000000

Understanding the data

Imbalanced data

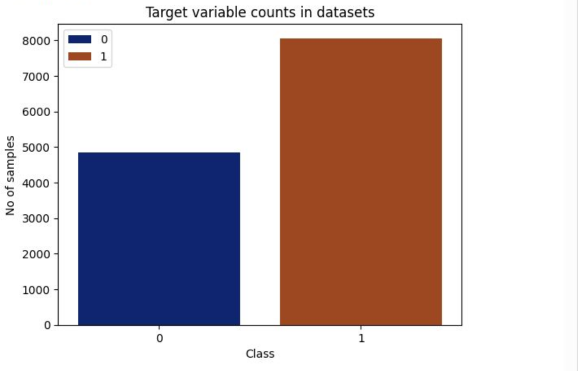
- Not Diagnosed within 90 days[0]: 4846
- Diagnosed within 90 day[1]: 8060
- Will be handled by resampling

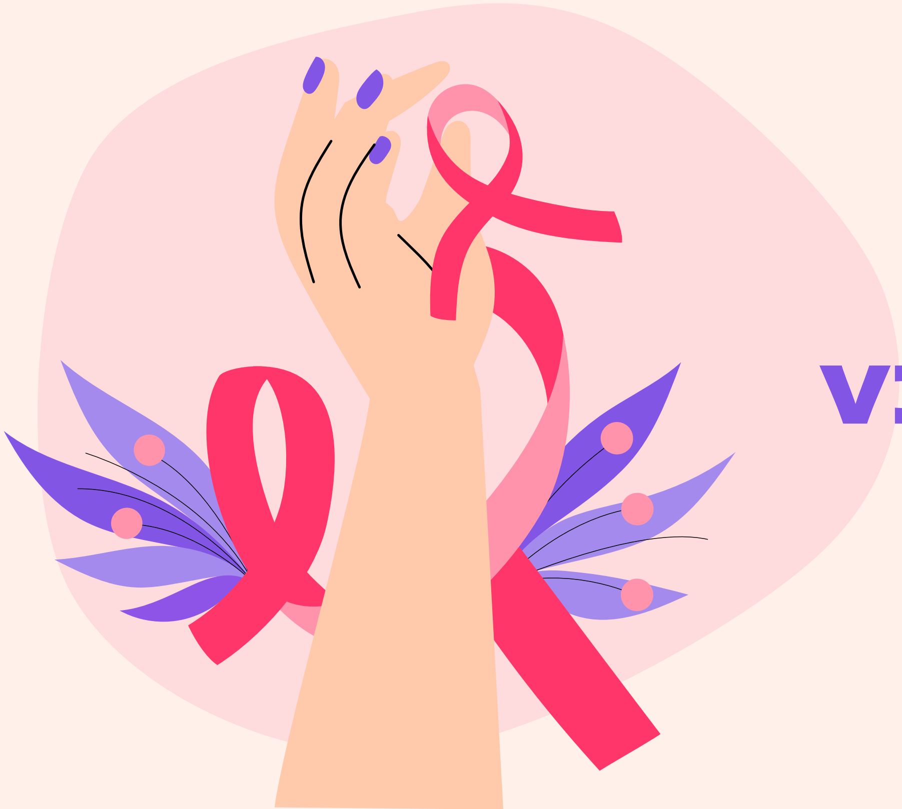
```python
target_column_name = 'DiagPeriodL90D'
target_variable = df_encoded[target_column_name]

# Addingthe target variable to the selected_df
selected_df[target_column_name] = target_variable
```

```python
### Checking and plotting for the class in the target variable
(unique, counts) = np.unique(selected_df['DiagPeriodL90D'], return_counts=True)
print(unique, counts)
sns.barplot(x=unique, y=counts, hue=unique, palette='dark', legend=True)
plt.xlabel("Class")
plt.ylabel("No of samples")
plt.xticks()
plt.title("Target variable counts in datasets")
plt.show()
plt.close()
```

```
[0 1] [4846 8060]
```

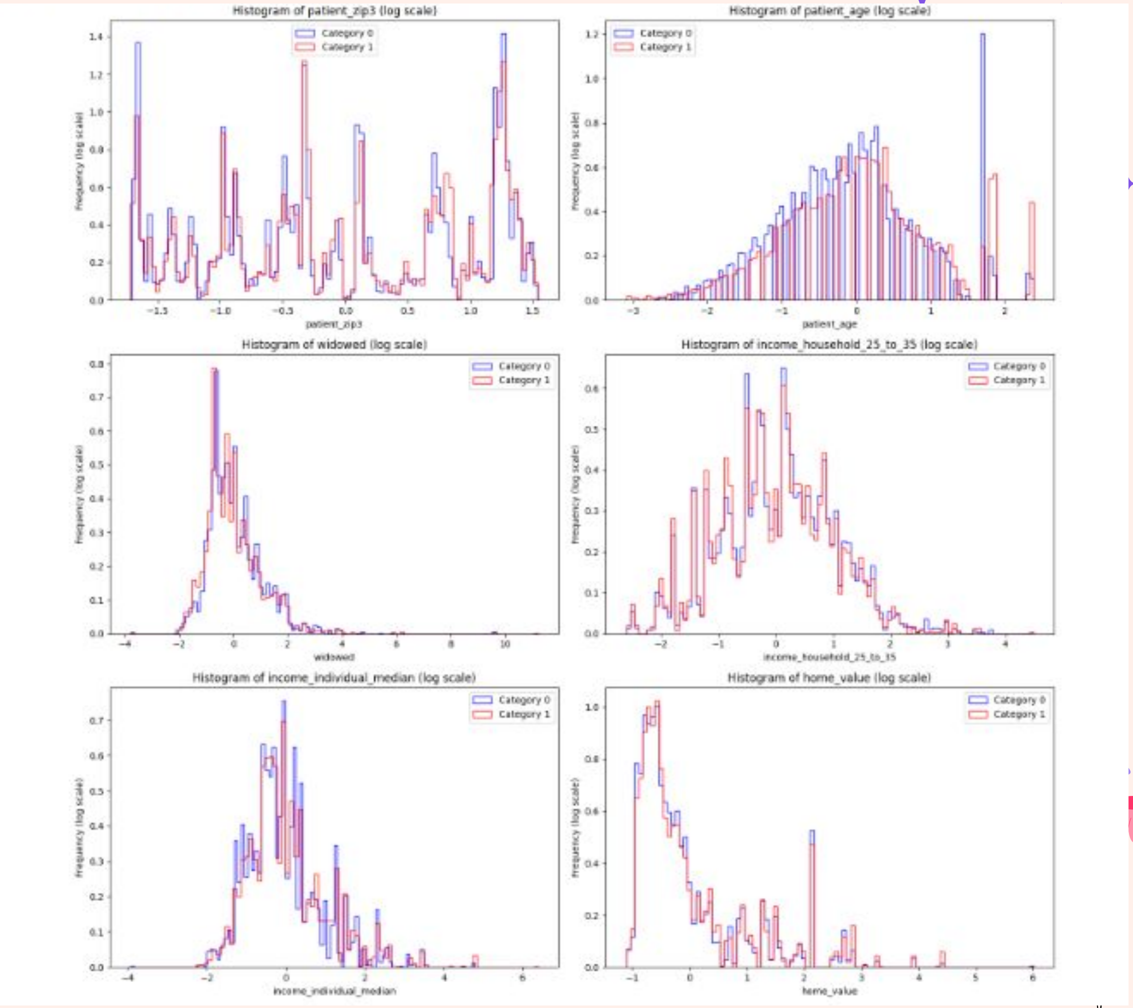


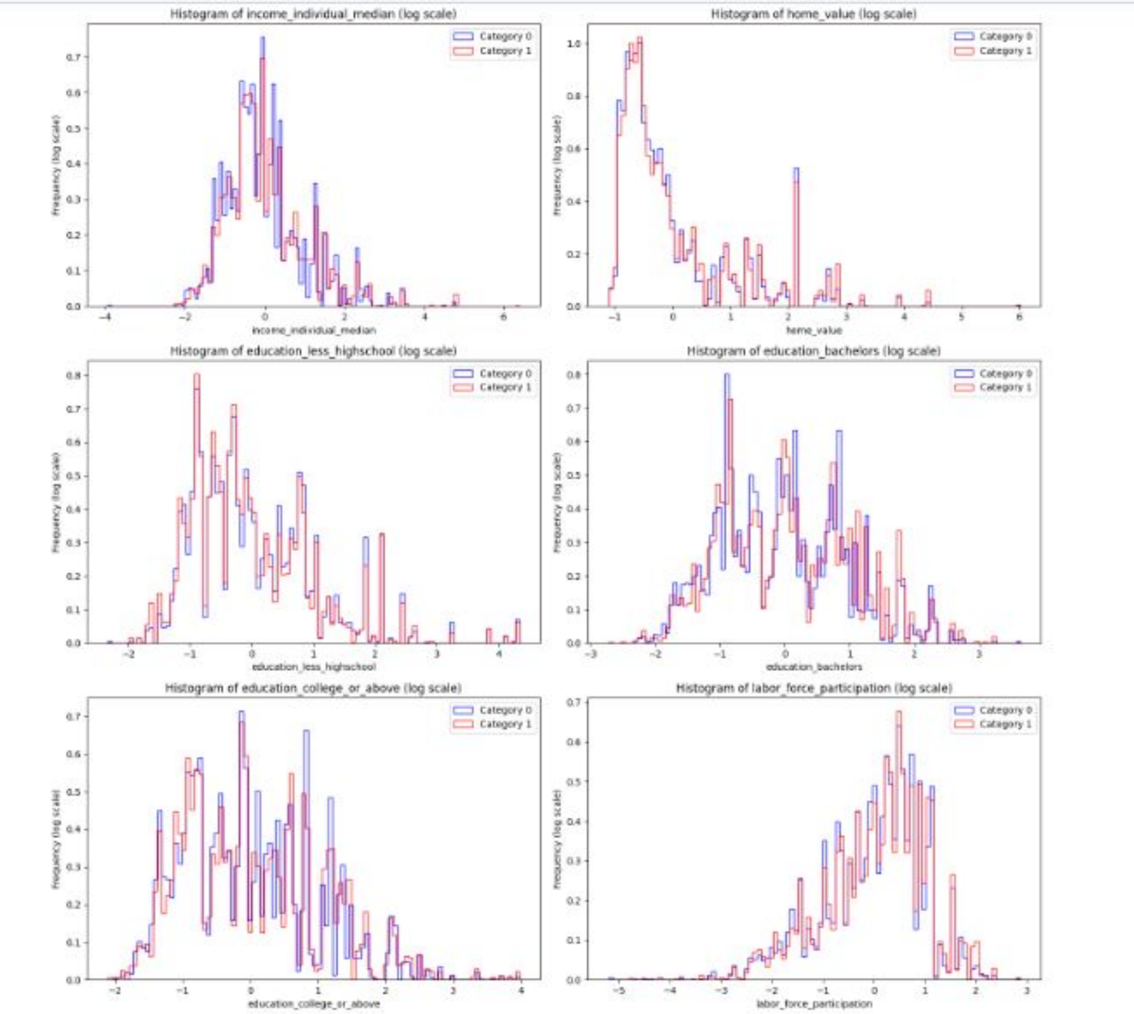


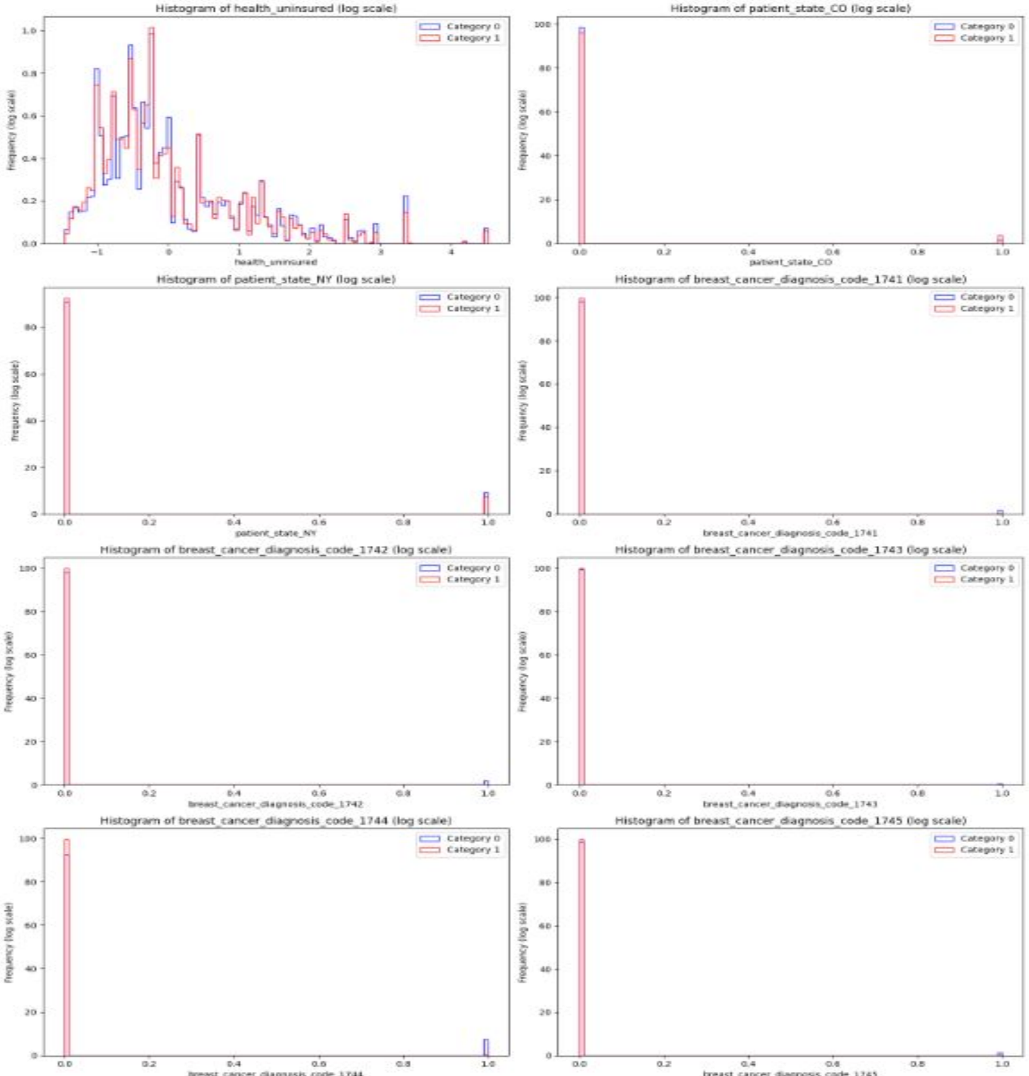
DATA VISUALIZATION

- Numerical and categorical datasets

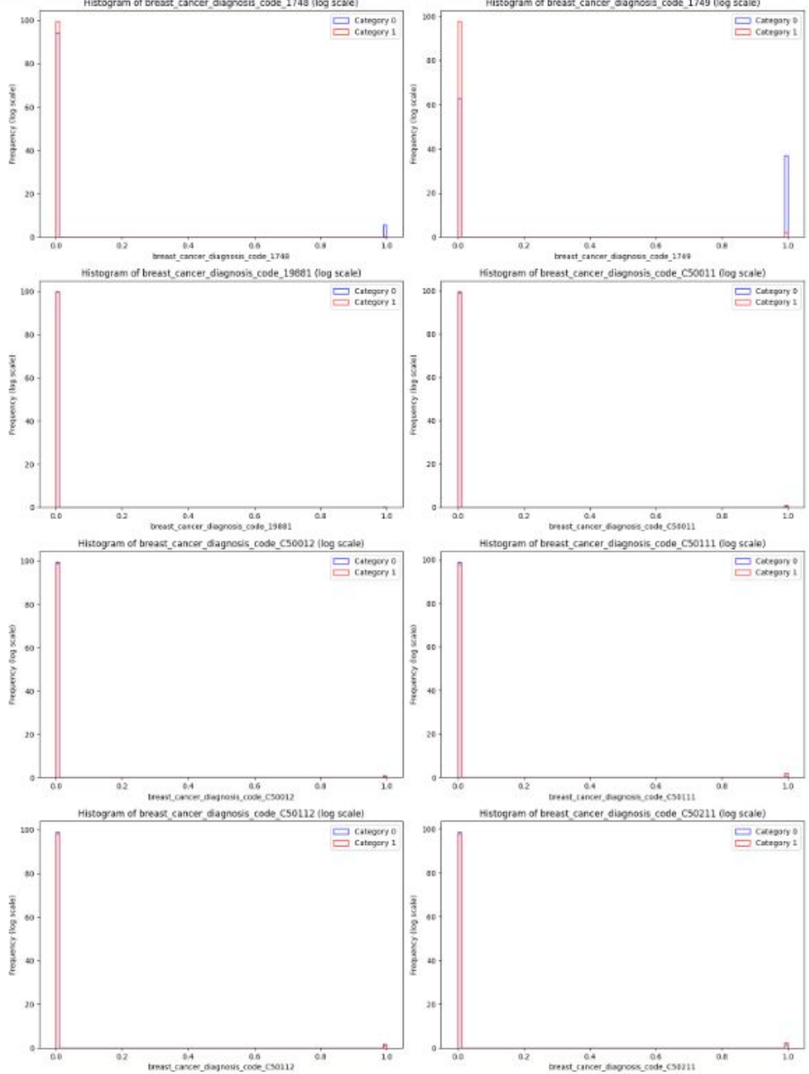
- Theses Histogram is showcasing the distribution of the target variable in each features

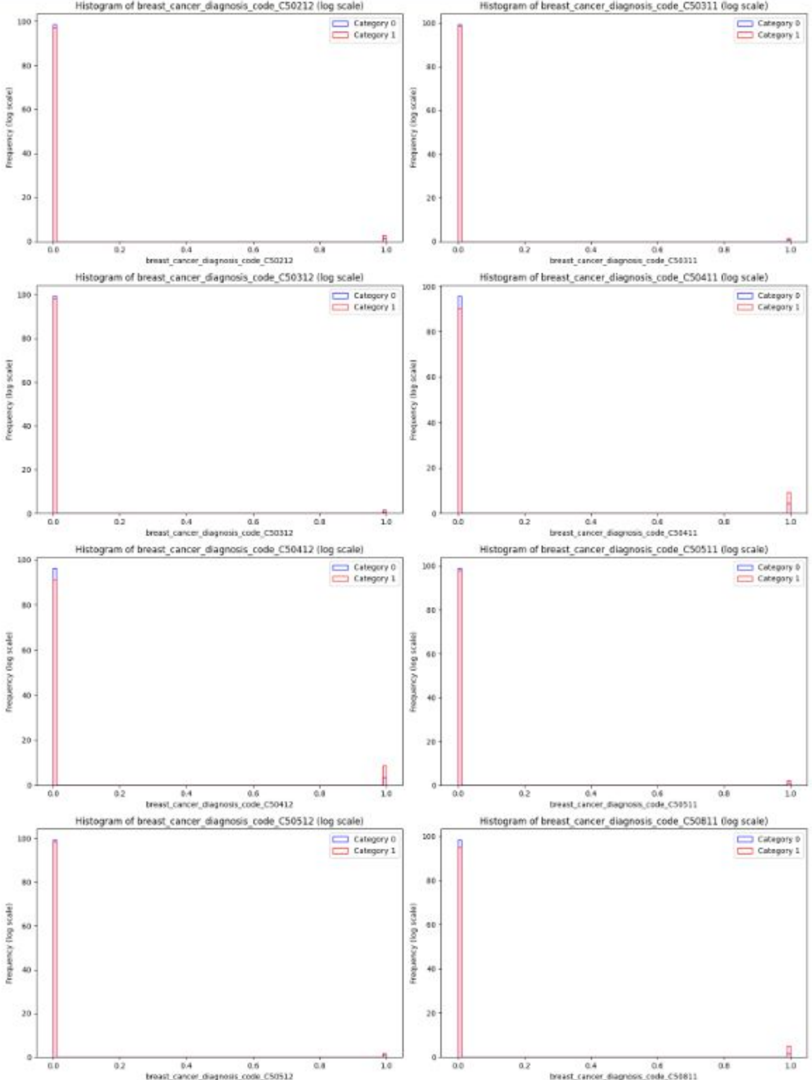


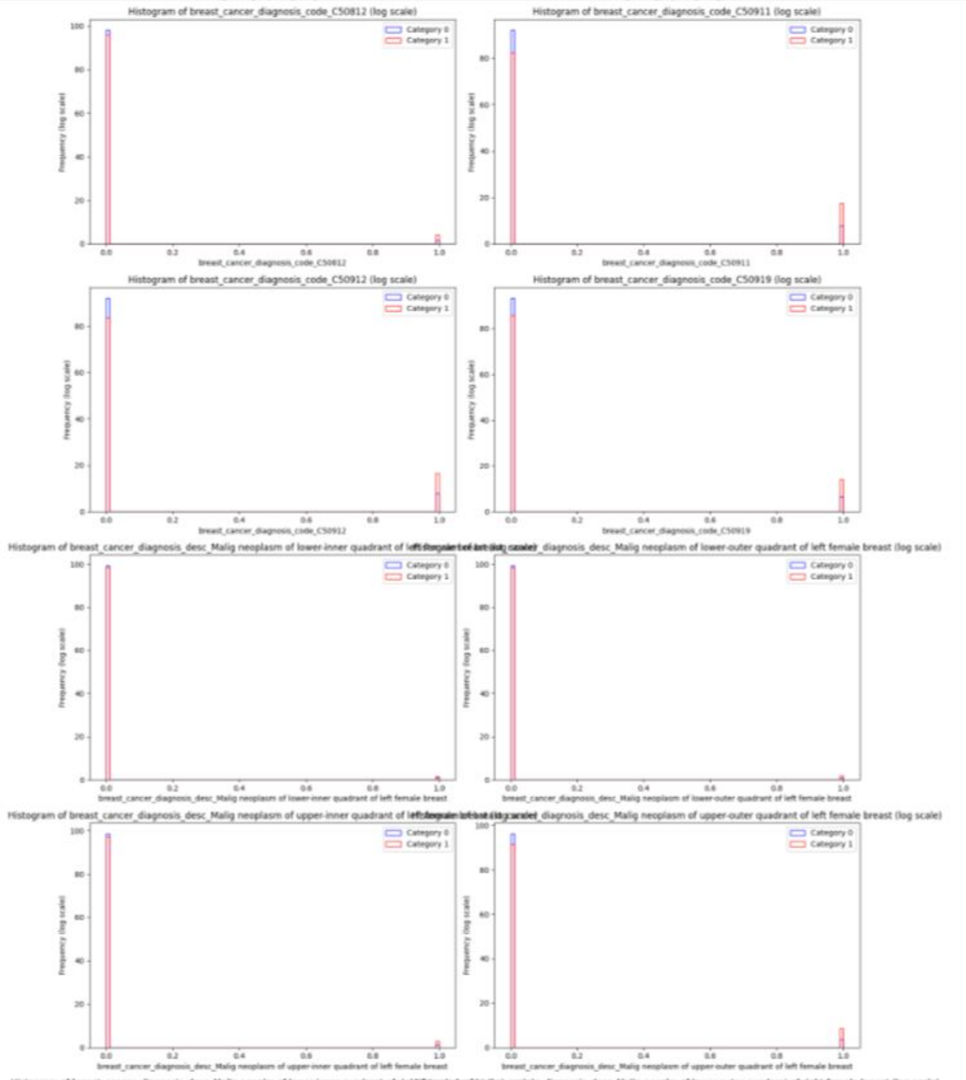

Histogram of income_individual_median (log scale)
Histogram of home_value (log scale)
Histogram of education_less_highschool (log scale)
Histogram of education_bachelors (log scale)
Histogram of education_college_or_above (log scale)
Histogram of labor_force_participation (log scale)
Category 0
Category 1
Frequency (log scale)
income_individual_median
home_value
education_less_highschool
education_bachelors
education_college_or_above
labor_force_participation



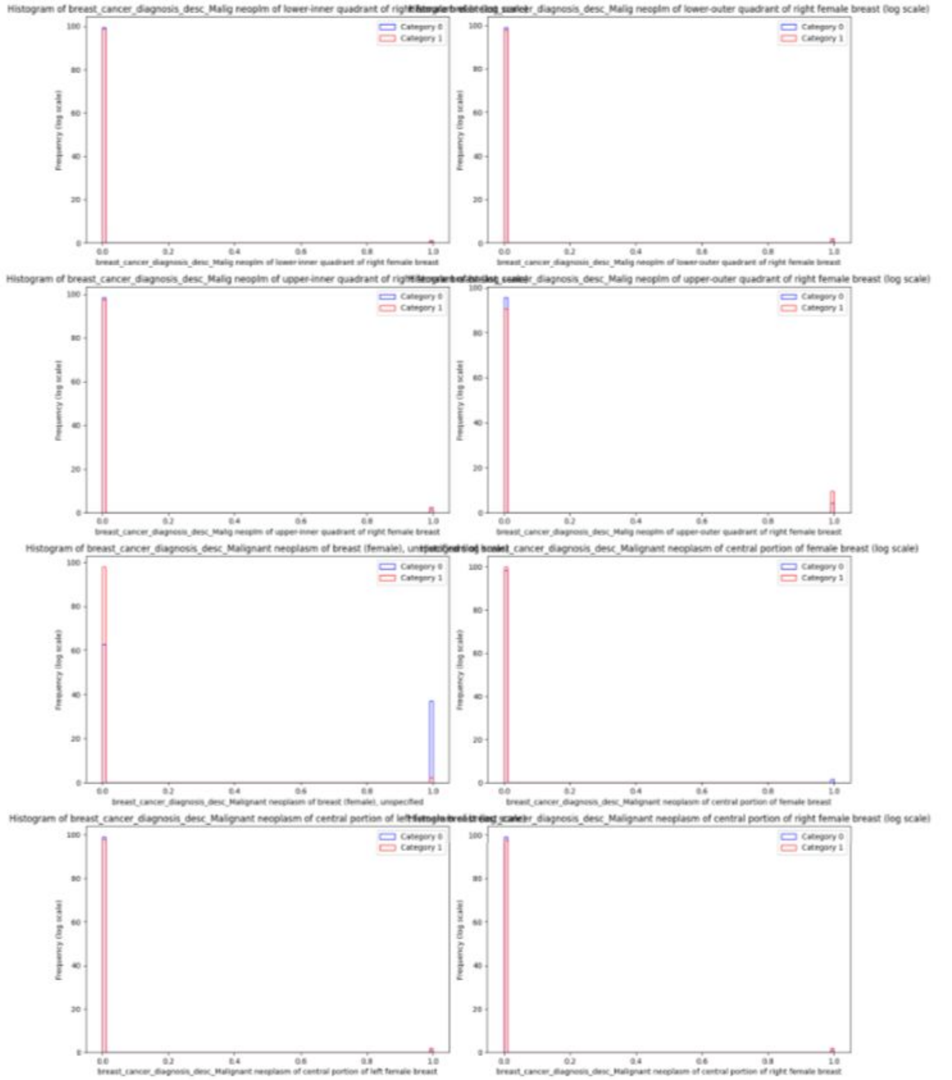
Histogram of health_uninsured (log scale)
Category 0
Category 1
Frequency (log scale)
health_uninsured
Histogram of patient_state_CO (log scale)
patient_state_CO
Histogram of patient_state_NY (log scale)
patient_state_NY
Histogram of breast_cancer_diagnosis_code_1741 (log scale)
breast_cancer_diagnosis_code_1741
Histogram of breast_cancer_diagnosis_code_1742 (log scale)
breast_cancer_diagnosis_code_1742
Histogram of breast_cancer_diagnosis_code_1743 (log scale)
breast_cancer_diagnosis_code_1743
Histogram of breast_cancer_diagnosis_code_1744 (log scale)
Histogram of breast_cancer_diagnosis_code_1745 (log scale)


Histogram of breast_cancer_diagnosis_code_1748 (log scale)
Histogram of breast_cancer_diagnosis_code_1749 (log scale)
Histogram of breast_cancer_diagnosis_code_19881 (log scale)
Histogram of breast_cancer_diagnosis_code_C50011 (log scale)
Histogram of breast_cancer_diagnosis_code_C50012 (log scale)
Histogram of breast_cancer_diagnosis_code_C50111 (log scale)
Histogram of breast_cancer_diagnosis_code_C50112 (log scale)
Histogram of breast_cancer_diagnosis_code_C50211 (log scale)
Category 0
Category 1
Frequency (log scale)

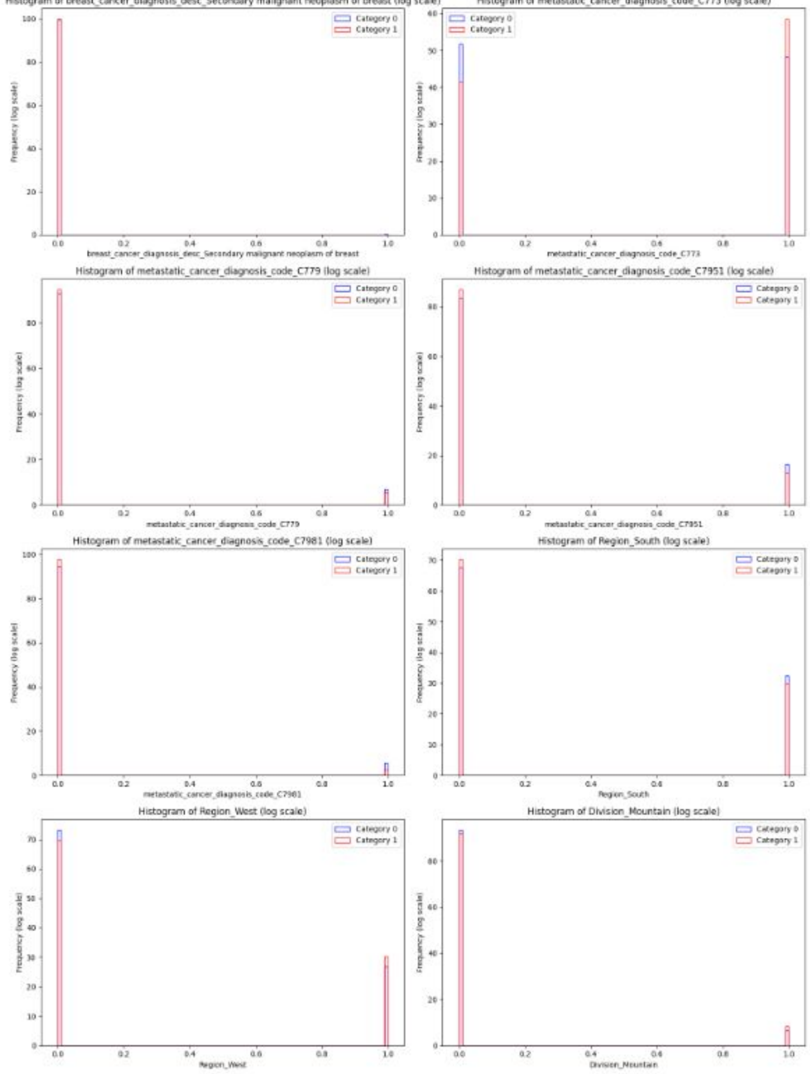

Histogram of breast_cancer_diagnosis_code_C50212 (log scale)
Histogram of breast_cancer_diagnosis_code_C50311 (log scale)
Histogram of breast_cancer_diagnosis_code_C50312 (log scale)
Histogram of breast_cancer_diagnosis_code_C50411 (log scale)
Histogram of breast_cancer_diagnosis_code_C50412 (log scale)
Histogram of breast_cancer_diagnosis_code_C50511 (log scale)
Histogram of breast_cancer_diagnosis_code_C50512 (log scale)
Histogram of breast_cancer_diagnosis_code_C50811 (log scale)
Category 0
Category 1
Frequency (log scale)
breast_cancer_diagnosis_code_C50212
breast_cancer_diagnosis_code_C50311
breast_cancer_diagnosis_code_C50312
breast_cancer_diagnosis_code_C50411
breast_cancer_diagnosis_code_C50412
breast_cancer_diagnosis_code_C50511
breast_cancer_diagnosis_code_C50512
breast_cancer_diagnosis_code_C50811



Histogram of breast_cancer_diagnosis_code_C50812 (log scale)
Category 0
Category 1
Frequency (log scale)
breast_cancer_diagnosis_code_C50812
Histogram of breast_cancer_diagnosis_code_C50911 (log scale)
breast_cancer_diagnosis_code_C50911
Histogram of breast_cancer_diagnosis_code_C50912 (log scale)
breast_cancer_diagnosis_code_C50912
Histogram of breast_cancer_diagnosis_code_C50919 (log scale)
breast_cancer_diagnosis_code_C50919
Histogram of breast_cancer_diagnosis_desc_Malig neoplasm of lower-inner quadrant of left female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malig neoplasm of lower-outer quadrant of left female breast (log scale)
breast_cancer_diagnosis_desc_Malig neoplasm of lower-inner quadrant of left female breast
breast_cancer_diagnosis_desc_Malig neoplasm of lower-outer quadrant of left female breast
Histogram of breast_cancer_diagnosis_desc_Malig neoplasm of upper-inner quadrant of left female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malig neoplasm of upper-outer quadrant of left female breast (log scale)
breast_cancer_diagnosis_desc_Malig neoplasm of upper-inner quadrant of left female breast
breast_cancer_diagnosis_desc_Malig neoplasm of upper-outer quadrant of left female breast



Histogram of breast_cancer_diagnosis_desc_Malig neoplm of lower-inner quadrant of right female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malig neoplm of lower-outer quadrant of right female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malig neoplm of upper-inner quadrant of right female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malig neoplm of upper-outer quadrant of right female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malignant neoplasm of breast (female), unspecified (log scale)
Histogram of breast_cancer_diagnosis_desc_Malignant neoplasm of central portion of female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malignant neoplasm of central portion of left female breast (log scale)
Histogram of breast_cancer_diagnosis_desc_Malignant neoplasm of central portion of right female breast (log scale)
Category 0
Category 1
Frequency (log scale)


Histogram of breast_cancer_diagnosis_desc_Secondary malignant neoplasm of breast (log scale)
Category 0
Category 1
Frequency (log scale)
breast_cancer_diagnosis_desc_Secondary malignant neoplasm of breast
Histogram of metastatic_cancer_diagnosis_code_C773 (log scale)
metastatic_cancer_diagnosis_code_C773
Histogram of metastatic_cancer_diagnosis_code_C779 (log scale)
metastatic_cancer_diagnosis_code_C779
Histogram of metastatic_cancer_diagnosis_code_C7951 (log scale)
metastatic_cancer_diagnosis_code_C7951
Histogram of metastatic_cancer_diagnosis_code_C7981 (log scale)
metastatic_cancer_diagnosis_code_C7981
Histogram of Region_South (log scale)
Region_South
Histogram of Region_West (log scale)
Region_West
Histogram of Division_Mountain (log scale)
Division_Mountain

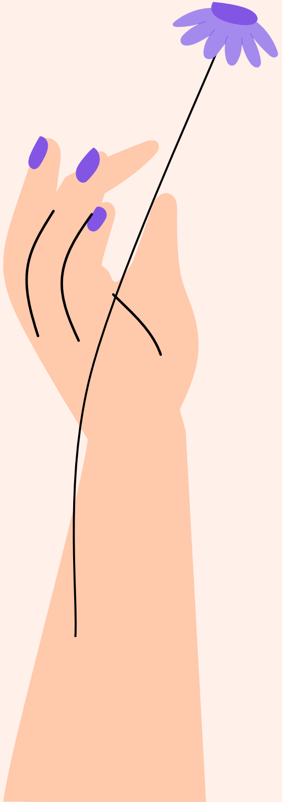
Data Preprocessing

DATA PREPROCESSING

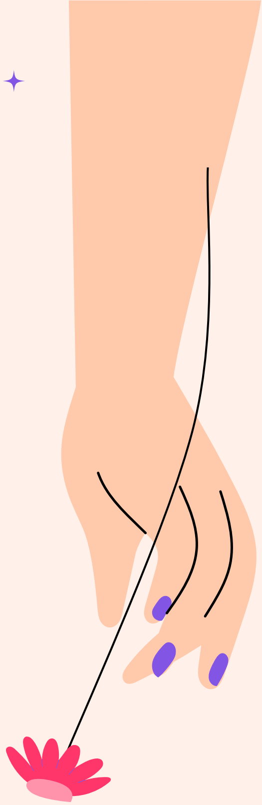
- Missing Data Handled
- Outliers handled with log transformation
- One hot encoded the categorical variables
- Dropped unnecessary columns like “ID”

DATA PREPROCESSING

- Using variance Thresholding for feature selections
- Repeated the same steps for the testing data



Model Building



Logistics Regression Model

Strengths

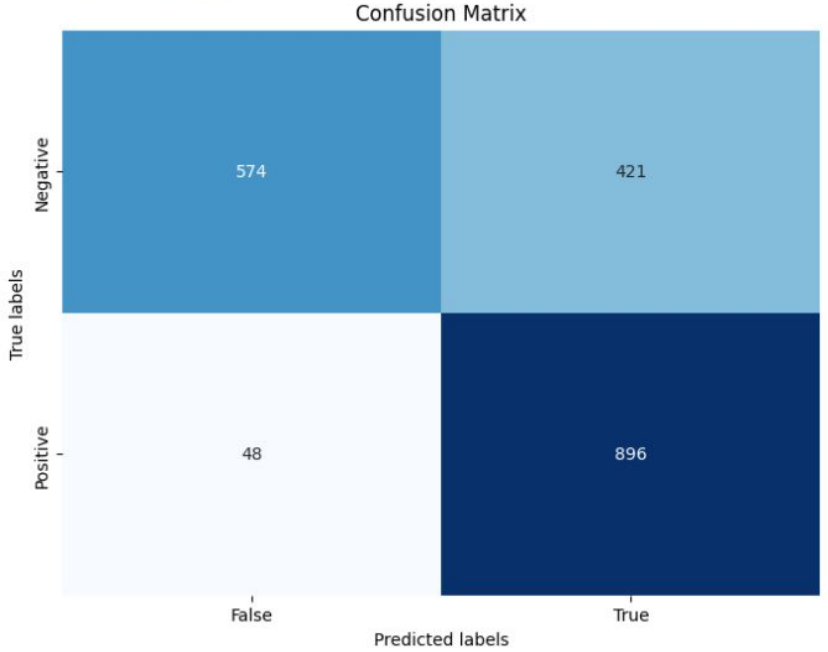
- Works well with both numerical and categorical data.
- Can handle large dataset with low computational cost
- Less susceptible to overfitting
- Provides easily interpretable results.

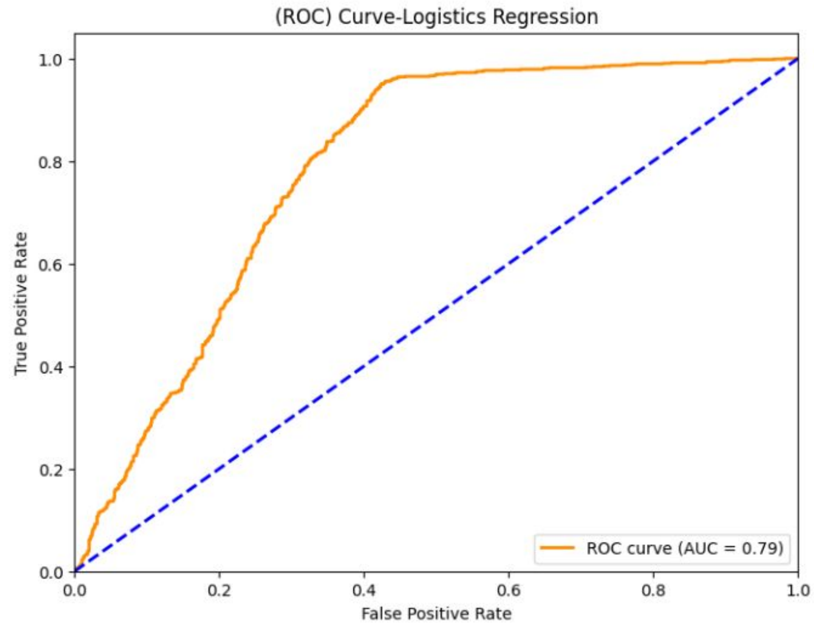
Weaknesses

- May not capture complex relationship within the data.
- Sensitive to multicollinearity
- Assumes a linear relationship

Confusion Matrix & ROC Curve for Logistic Regression

```
Accuracy: 0.7581227436823105
Precision: 0.680334092634776
Recall: 0.9491525423728814
F1-score: 0.7925696594427245
```





Gradient Boosting Model

Strengths

- Achieve high accuracy
- Handles complex relationships
- Helps Identify feature that contribute the most to the model's predictions
- Robust to Outliers

Weaknesses

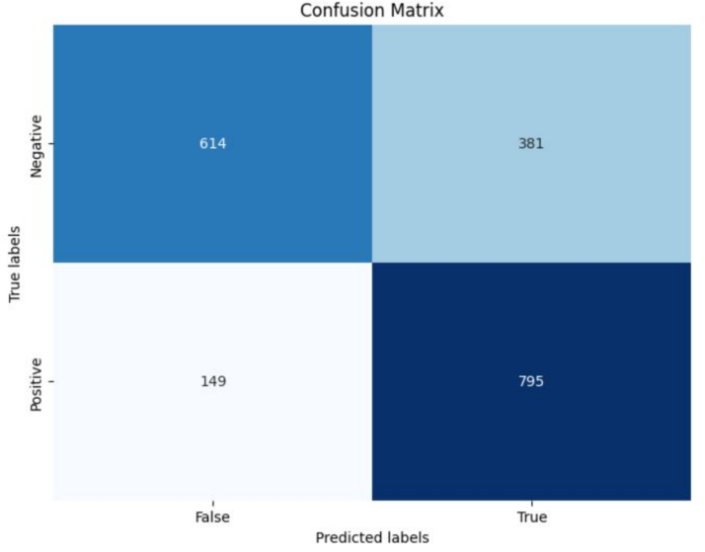
- Computationally expensive.
- Prone to overfitting
- Requires careful overfitting of Hyperparameters

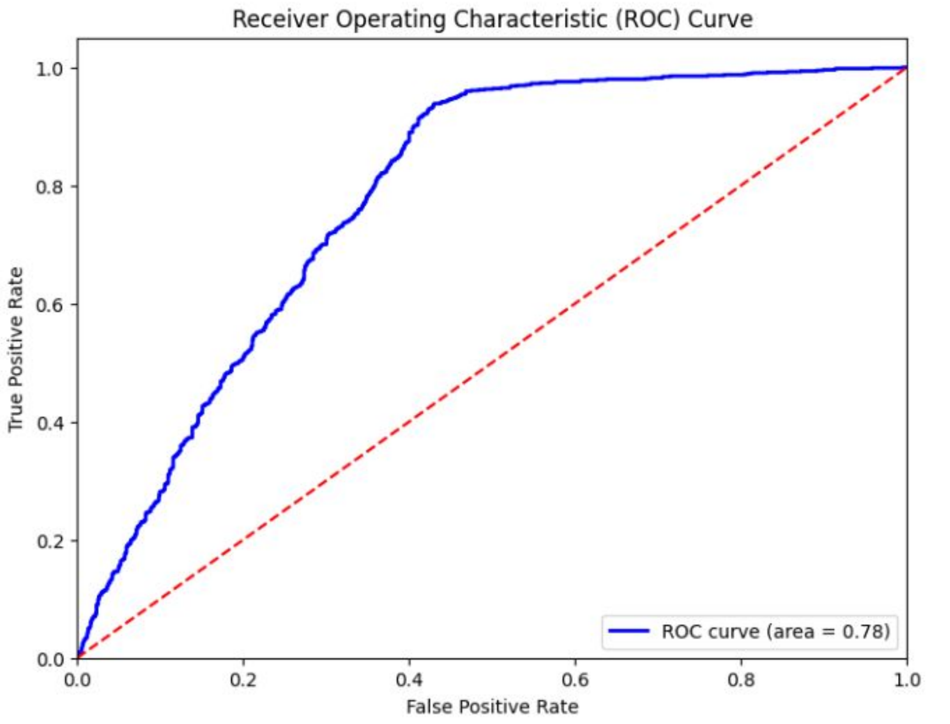
Confusion Matrix & ROC Curve for XGBoost

```
Accuracy: 0.7266632284682826
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.62      0.70       995
           1       0.68      0.84      0.75       944

    accuracy                           0.73      1939
   macro avg       0.74      0.73      0.72      1939
weighted avg       0.74      0.73      0.72      1939
```





Deep Learning

FNN Model

Strengths

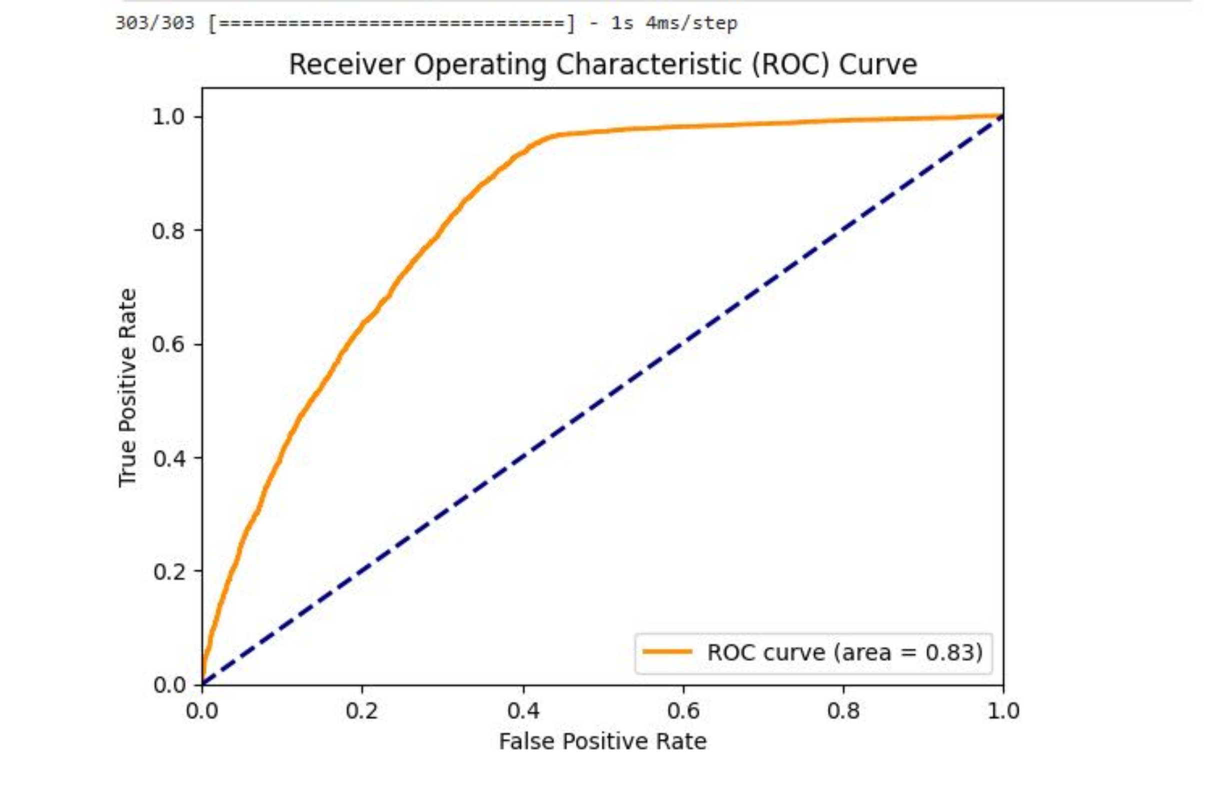
- Can deal with non-linearity in a dataset
- Can handle numerical and categorical dataset
- Can handle large datasets

Weaknesses

- Computationally expensive.
- Prone to overfitting

ROC Curve of the FNN model

- **FNN Model**
- **10 Epochs**



MODEL EVALUATION

Logistics Regression

```
Accuracy: 0.7581227436823105
Precision: 0.680334092634776
Recall: 0.9491525423728814
F1-score: 0.7925696594427245
```

XGBoost

```
Accuracy: 0.7266632284682826
Classification Report:
              precision    recall  f1-score

           0       0.80      0.62      0.70
           1       0.68      0.84      0.75
```

FNN Model

```
Accuracy: 0.7407140135765076
Precision: 0.7372381687164307
Recall: 0.7480396032333374
```

Kaggle Score of Each Model on the Test data

submission_lr.csv	Complete (after deadline) · 14...	0.718
submission_xgb1.csv	Complete (after deadline) · 15...	0.733
submission_fnn.csv	Complete (after deadline) · 14...	0.787

MODEL EVALUATION

- All of my model performed well, all their accuracy ranging in 70-80%, though the best model out of all of them was my deep learning model with a score of 0.74 on the training data & 0.78 on the test data.

- Though my model performed well, It definitely can be improved. I might be able to achieve that by scaling my datasets, and selecting more features for my models.


THE END