Final Project

Supervised Machine Learning on Adult Dataset

Contents

1 Project Topic

Task

Project task is to develop and create binary classification models.

Goals

- Primary
- Secondary

Census Income Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	750875

2 Data

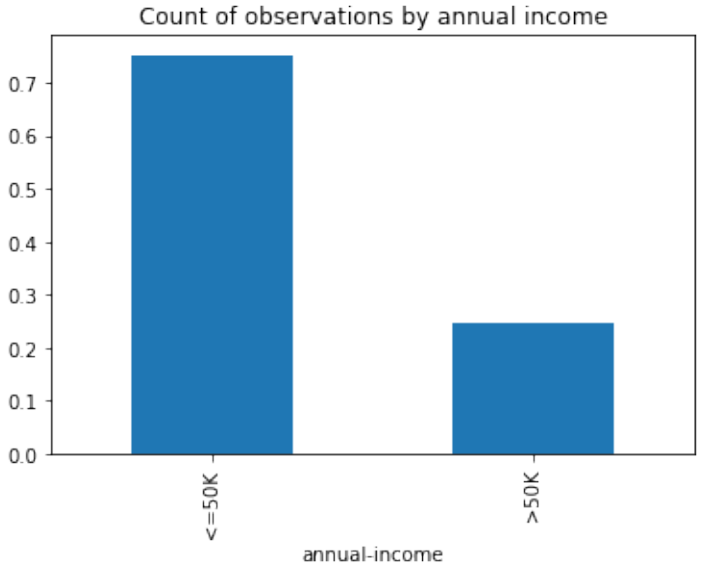
Data Source
https://archive.ics.uci.edu/ml/datasets/Census+Income

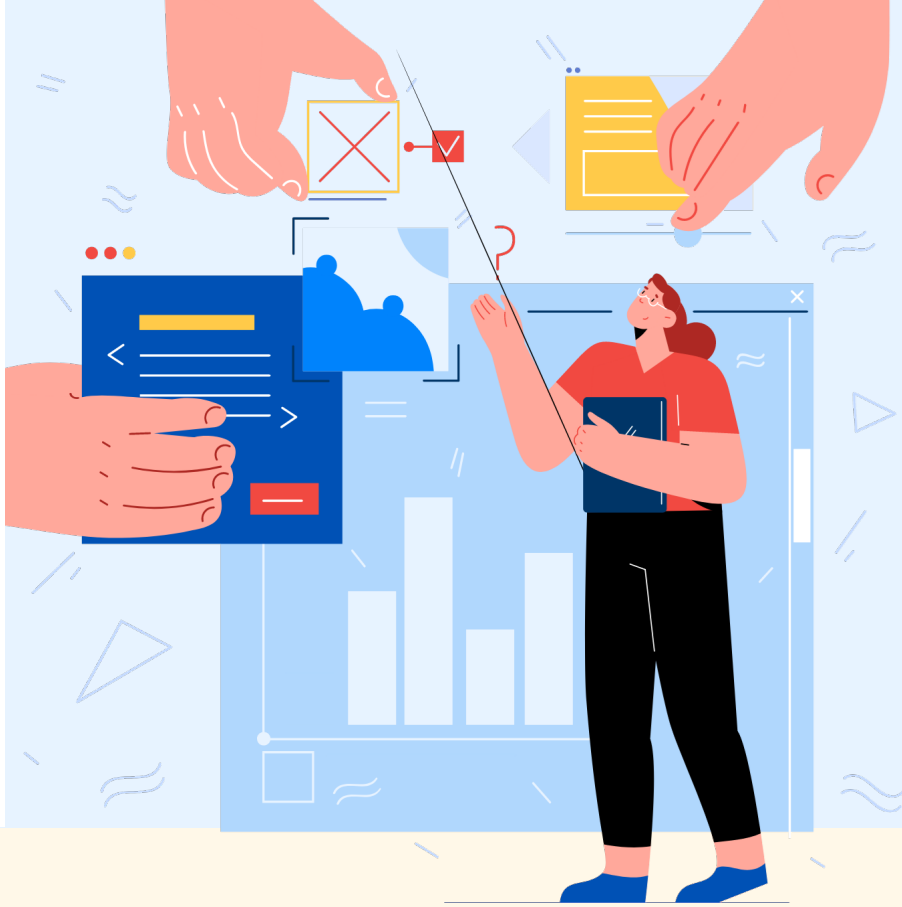
Data Description
Characteristics of dataset.

Data Attributes
Information of attributes.

Data Summary
Checking the summary of data.

❸ Data Cleaning





Missing Values

na_values='?' is the argument used while reading in dataset.

Check of Imbalanced Data

We will first clean target variable.

Conclusion

4

Exploratory Data Analysis

* Refer to notebook for EDA and visualizations.

Multicollinearity?

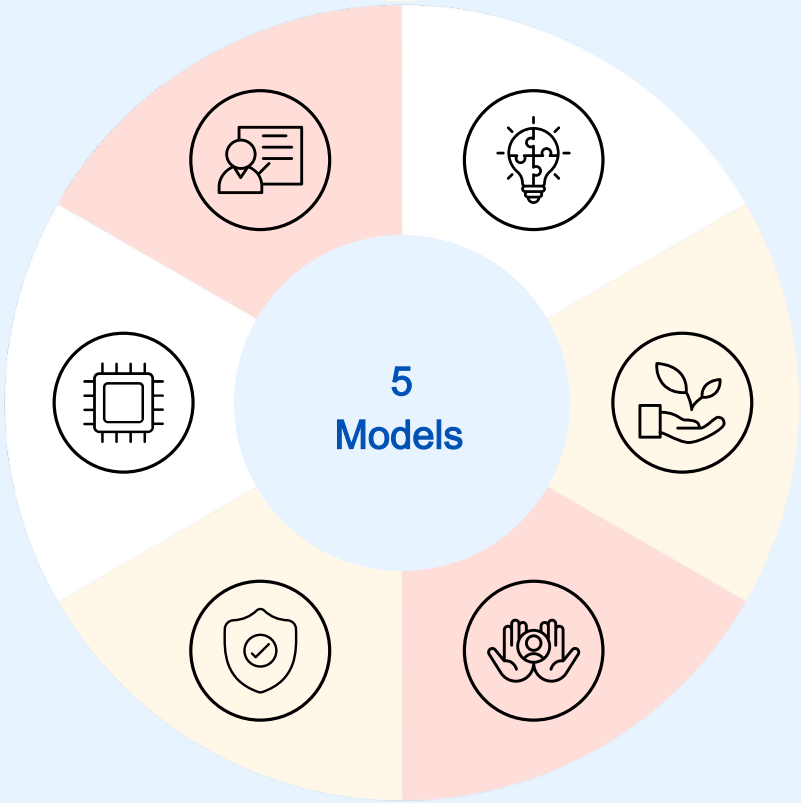
We check for dependency between marital status and relationship.

Encoding Categorical Variables

We represent annual income of >50K as 1 and <=50K as 0.

Standardizing Feature Set

We standardize the age, hours-per-week and



Multicollinearity Test

It’s best if you keep your text within 3-5 lines. Less is more!

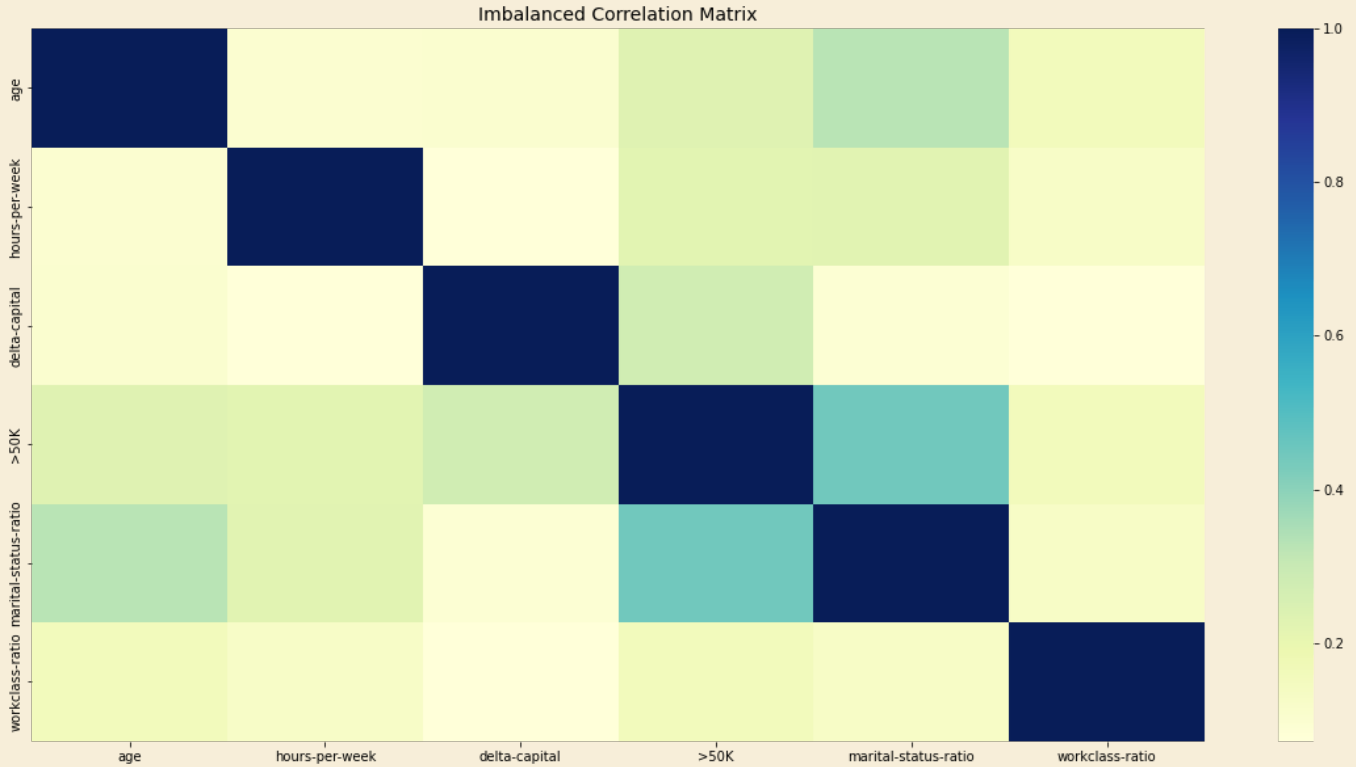
Synthetic Minority Oversampling Technique

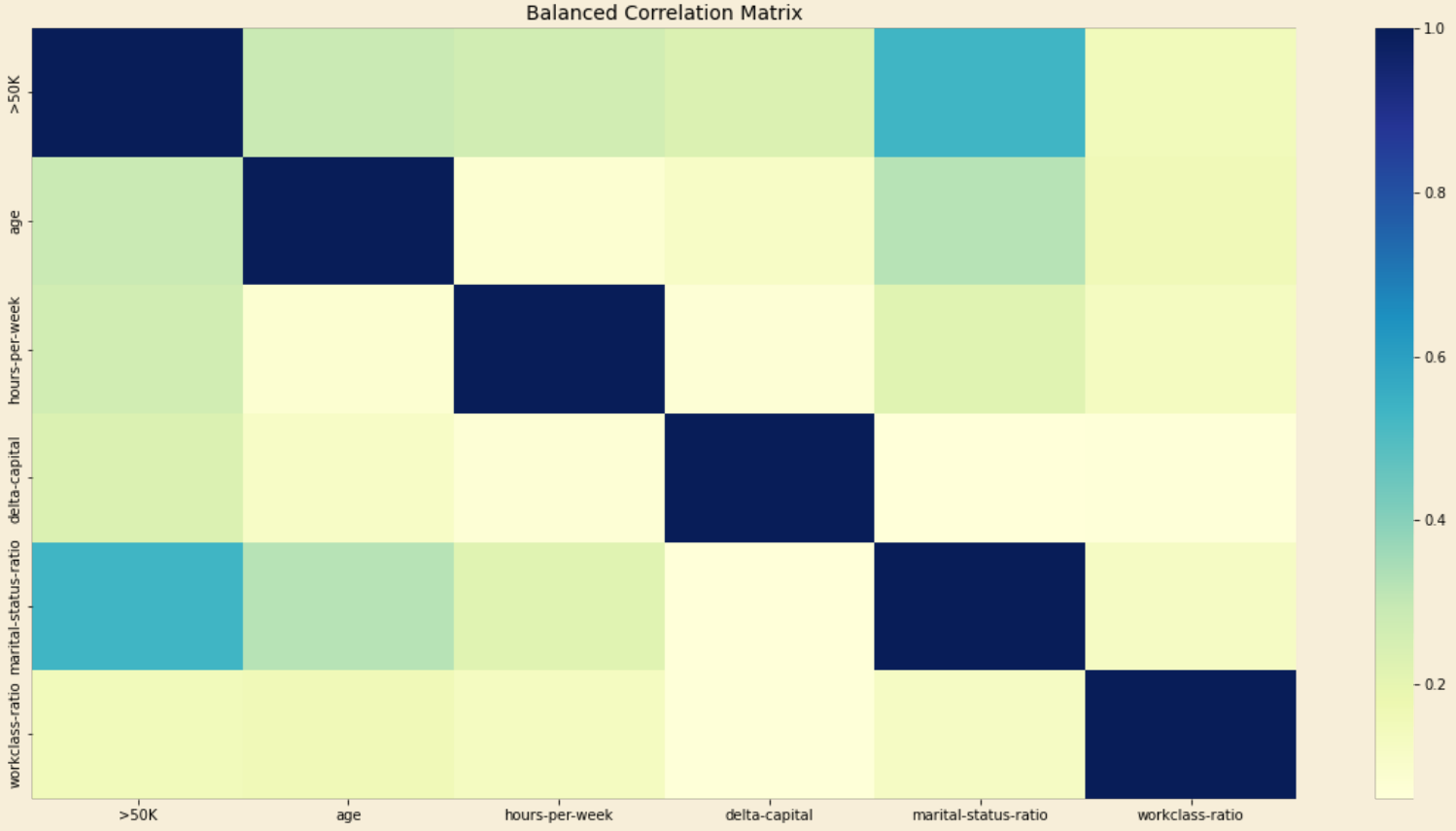
We use SMOTE to balance our dataset by over-sampling the minority class.

Train-Test Split

We will divide the dataset into a 70:30 ratio.

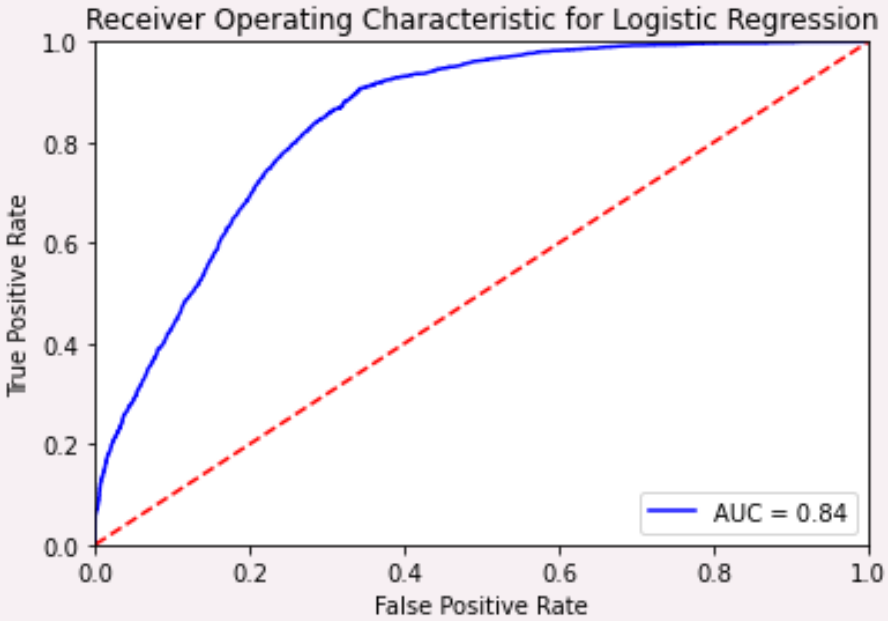
SMOTE: Synthetic Minority Oversampling Technique



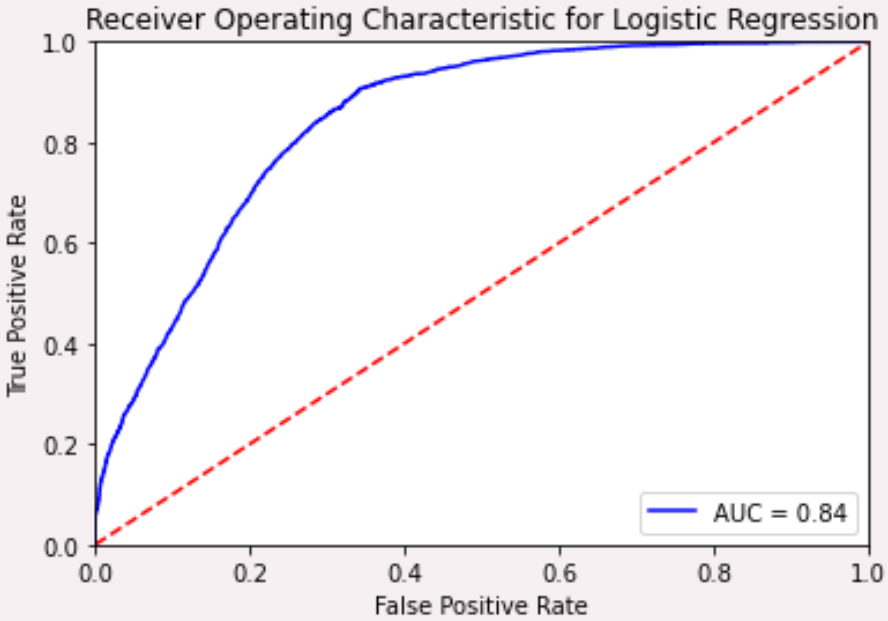

Balanced Correlation Matrix
>50K
age
hours-per-week
delta-capital
marital-status-ratio
workclass-ratio
1.0
0.8
0.6
0.4
0.2

Evaluation Metric

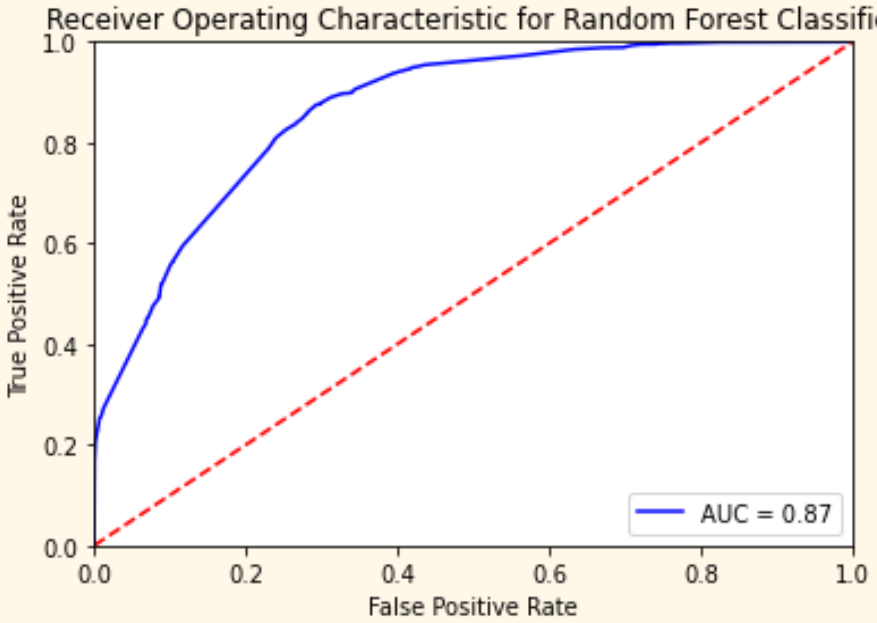
- Logistic Regression



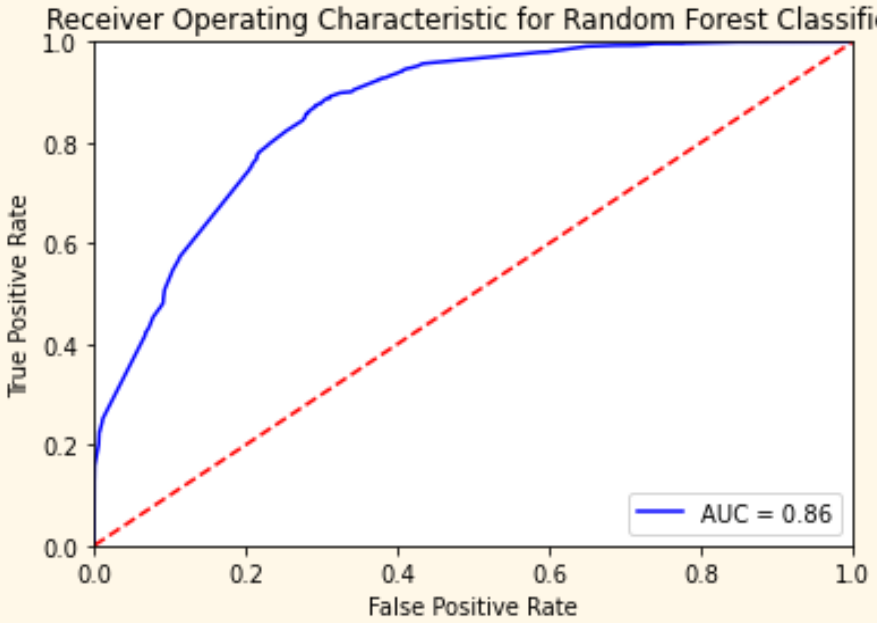
- Logistic Regression (with SMOTE)



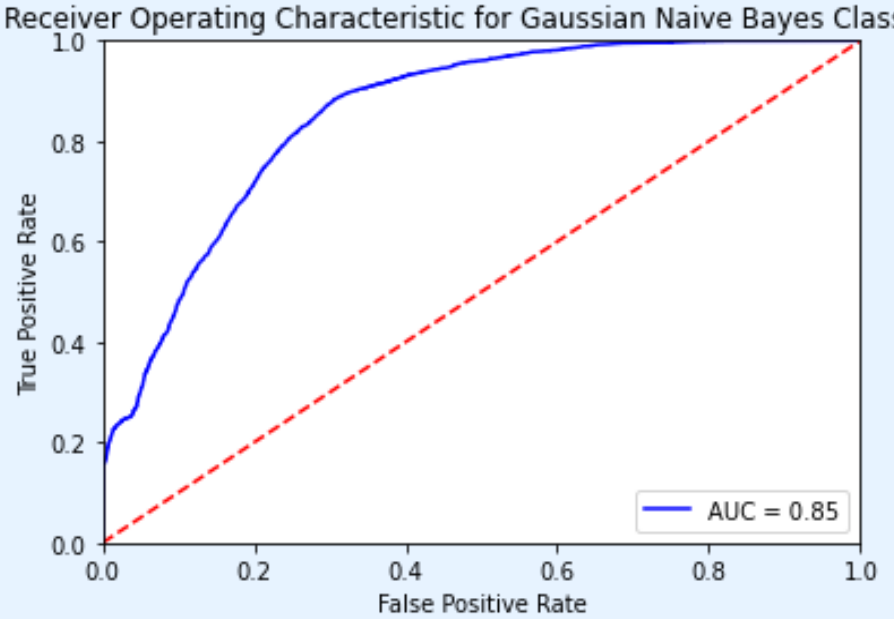
- Random Forest Classifier



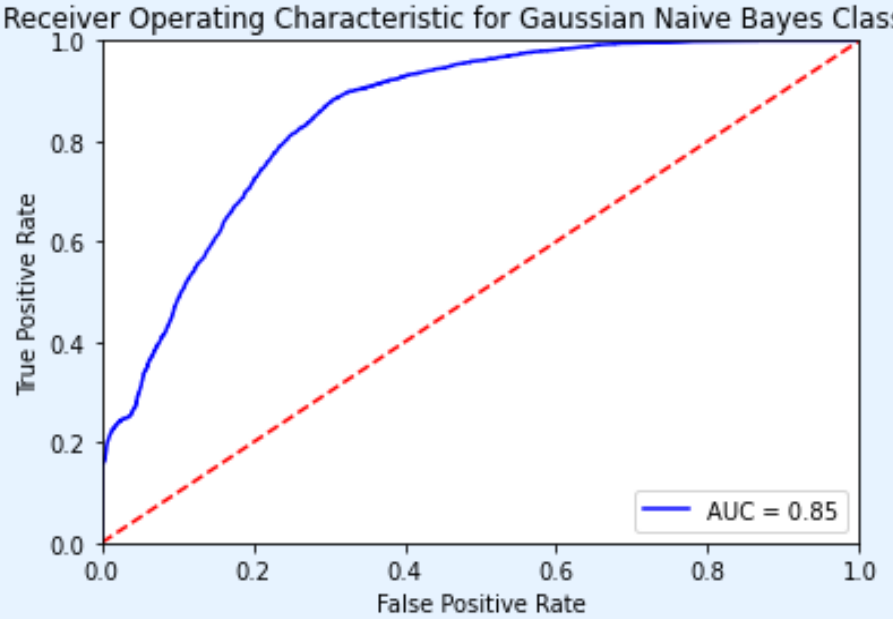
- Random Forest Classifier (SMOTE)



- Gaussian Naive Bayes



- Gaussian Naive Bayes (SMOTE)



Results and Analysis

Random Forest Classifier

Without SMOTE

Random Forest Train accuracy 0.79 Random Forest Test accuracy 0.79

	precision	recall	f1-score	support
0	0.79	1.00	0.88	10194
1	0.99	0.17	0.30	3304

With SMOTE

Random Forest Train accuracy 0.78 Random Forest Test accuracy 0.74

	precision	recall	f1-score	support
0	0.95	0.70	0.81	10194
1	0.49	0.88	0.63	3304

AUC

The AUC score for Random Forest Classifier is 0.87 for without SMOTE and is 0.86 for with SMOTE.

Logistic Regression

Without SMOTE

Logistic Regression Train accuracy 0.72 Logistic Regression Test accuracy 0.72

	precision	recall	f1-score	support
0	0.94	0.68	0.79	10194
1	0.47	0.87	0.61	3304

With SMOTE

Logistic Regression Train accuracy 0.77 Logistic Regression Test accuracy 0.72

	precision	recall	f1-score	support
0	0.94	0.68	0.79	10194
1	0.47	0.87	0.61	3304

AUC

The AUC score for Logistic Regression is 0.84 in both cases - with or without SMOTE.

Gaussian Naive Bayes

Without SMOTE

Gaussian Naive Bayes Train accuracy 0.79 Gaussian Naive Bayes Test accuracy 0.79

	precision	recall	f1-score	support
0	0.81	0.94	0.87	10194
1	0.66	0.34	0.44	3304

With SMOTE

Gaussian Naive Bayes Train accuracy 0.67 Gaussian Naive Bayes Test accuracy 0.79

	precision	recall	f1-score	support
0	0.84	0.91	0.87	10194
1	0.61	0.45	0.52	3304

AUC

The AUC score for Gaussian Naive Bayes is 0.85 in both cases - with or without SMOTE.

Discussion and Conclusion

- Data cleaning is crucial for machine learning pipeline.
- Exploratory Data Analysis is critical to identify the most significant factors.
- Statistical tests can easily identify the patterns and develop a solid feature set.
- Handling of imbalanced dataset is challenging.
- Accuracy may not be the best metric all the time.
- AUC score provides a better understanding for model performance when the dataset is imbalanced.

Reference: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.