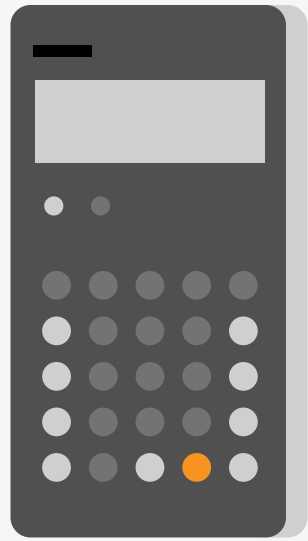
Global Financial Inclusion

General Assembly

Data Science DATR-501

Yaksh Birla

July 12, 2023



Project Overview

Today, over 76% of adults worldwide have an account at a financial institution or a mobile money account.[1] However, there is still large disparity in the level of financial inclusion across different geographies and demographics.

This project aims to address the ongoing global challenge of ensuring equitable access to financial services, empowering individuals and promoting economic growth through financial inclusion.

Overarching Question

Can we accurately predict how financially included an individual is, regardless of which country they are in?

1. Global Findex Report, World Bank Group. Source: https://www.worldbank.org/en/publication/globalfindex/interactive-executive-summary-visualization

Collecting and Merging the Data

Global Findex Database (from World Bank)

- Nationally representative surveys of ~128K adults in 123 economies
- Source of data on global access to financial services from payments to savings and borrowing with over 120 features

	economy	economycode	regionwb	pop_adult	wpid_random	wgt	female	age	educ
0	Afghanistan	AFG	South Asia	22647496.0	140343632	0.774286	1	19.0	2
1	Afghanistan	AFG	South Asia	22647496.0	167823412	0.766367	1	40.0	1
2	Afghanistan	AFG	South Asia	22647496.0	182483450	0.588983	1	25.0	1
3	Afghanistan	AFG	South Asia	22647496.0	170778240	2.572345	1	40.0	1
4	Afghanistan	AFG	South Asia	22647496.0	170712642	0.525471	2	27.0	3

Human Development Index (from the UN)

- A multidimensional weighted index that accounts for health, education and standard of living across countries
- Incorporate country-specific development factors for over 180 countries

	economycode	economy	hdi_rank_2021	hdi_2020	hdi_2021
0	AFG	Afghanistan	180.0	0.483	0.478
1	AGO	Angola	148.0	0.590	0.586
2	ALB	Albania	67.0	0.794	0.796
3	AND	Andorra	40.0	0.848	0.858
4	ARE	United Arab Emirates	26.0	0.912	0.911

Using economycode as a common identifier, we can merge both datasets:

```
# Create new variable for merged datasets on economycode
merged_data = pd.merge(micro_world, hdi_report, on="economycode", how="left")
```

Data Cleaning, Feature Importance and Visualization

Data Cleaning

- Economies (Countries) and Regions:
  - Remove rows with NAs in Regionwb column
  - Remove China due to different survey used
- Missing Age Data
  - Fill NAs with Median age (This has no effect on the model)

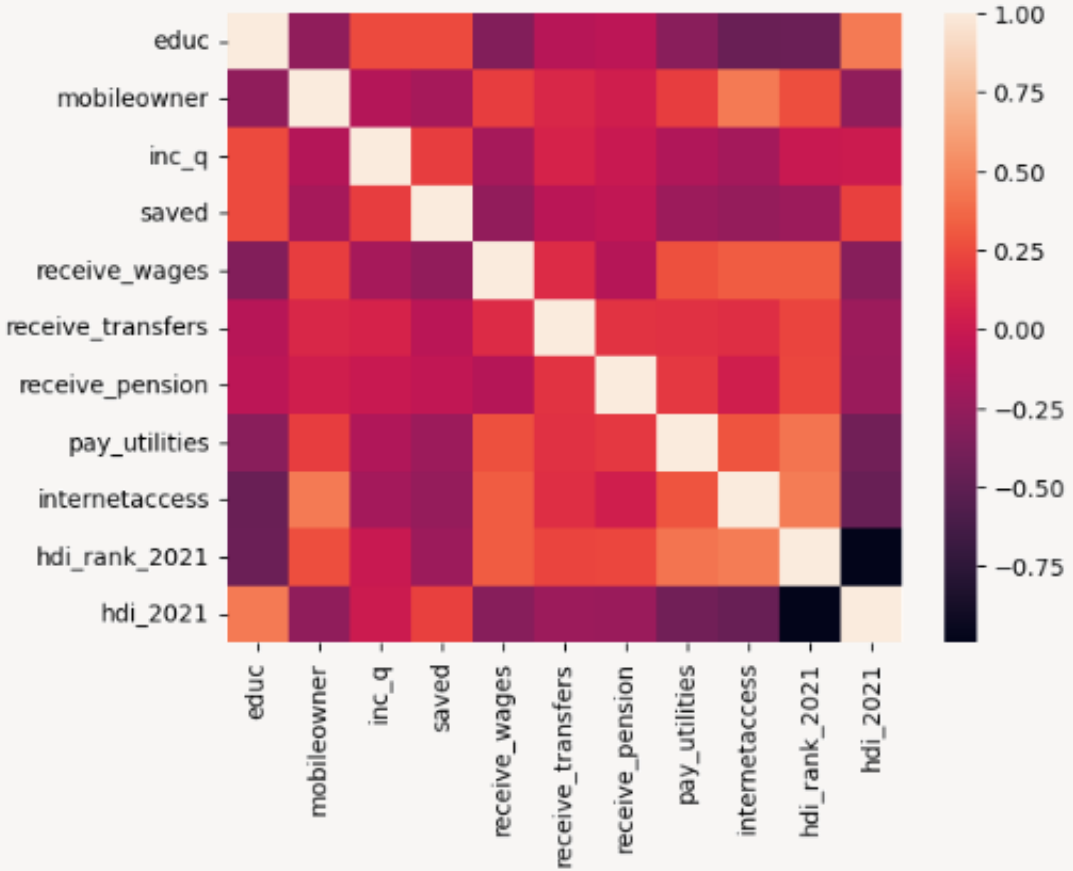
Remove Extraneous Values

- Education: Data dictionary showed that only responses 1, 2 and 3 are valid in the survey. However, the dataset also had instances where respondents had input ‘4’ and ‘5’ for Education. These were removed.

Combining Datasets

- Create new variable for merged datasets on economycode
- Remove null values of economycode on merged dataset
  - Identified Kosovo (Code: XKX) was found in HDI dataset but not in the Findex dataset

Correlation Matrix of Selected Features



Feature Selection and Pipeline Instantiation

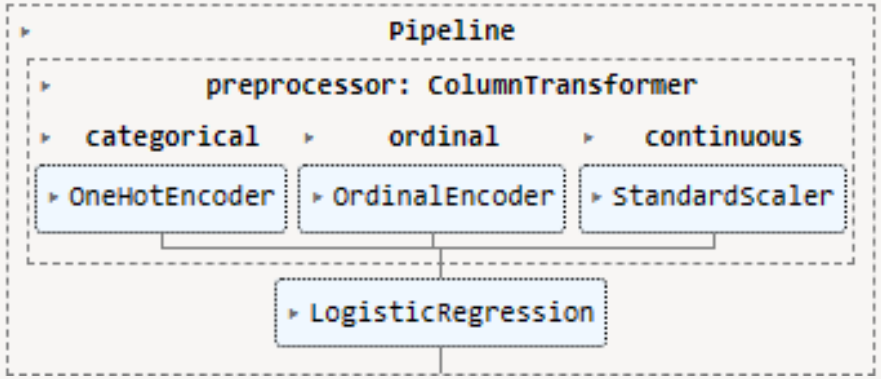
Global Findex Data

Features	Description	Variable Type
Educ	Level of education (=1, 2, or 3)	Ordinal
Mobileowner	Owns a phone (=1 to 4)	Categorical
Inc_q	Household income quintile (=1 to 5)	Ordinal
Saved	=1 if saved, 0 if not	Categorical
Receive_Wages	Wage payment received (=1 to 5)	Categorical
Receive_Transfers	Govt. transfer payment received (=1 to 5)	Categorical
Receive_Pensions	Pension received (=1 to 5)	Categorical
Pay_Utilities	Paid a utility bill (=1 to 5)	Categorical
InternetAccess	Internet access (=1 to 4)	Categorical

HDI Report

HDI_Rank_2021	Country HDI Rank (=1 to 180)	Ordinal
HDI_2021	Country HDI Score (=0 to 1)	Continuous

```
preprocessor = ColumnTransformer(
    transformers=[
        ('categorical', OneHotEncoder(), ['mobileowner',
'receive_wages', 'receive_transfers', 'receive_pension',
'saved', 'pay_utilities', 'internetaccess']),
        ('ordinal', OrdinalEncoder(), ['educ', 'inc_q',
'hdi_rank_2021']),
        ('continuous', StandardScaler(), ['hdi_2021'])])
```

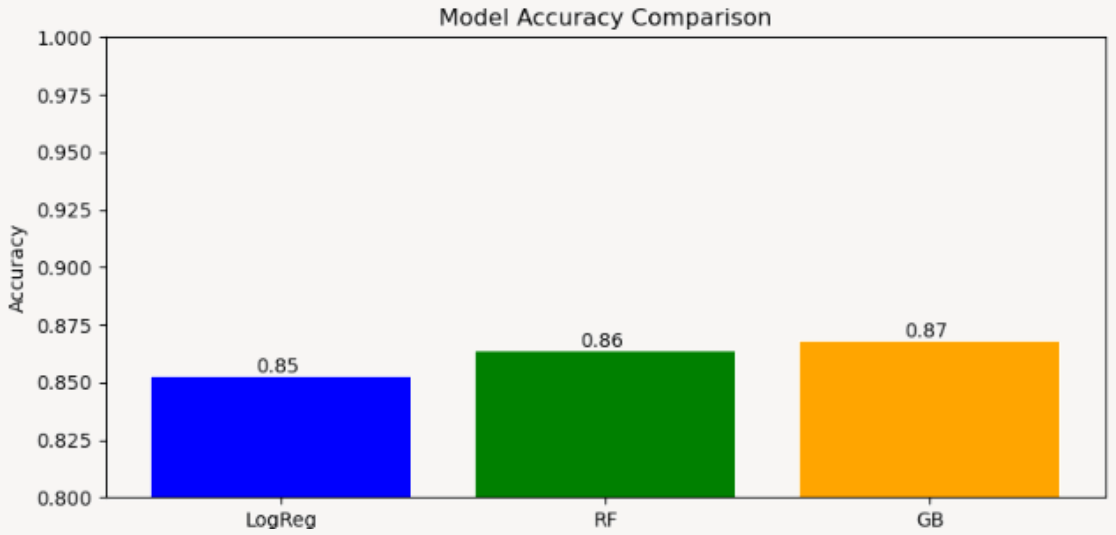


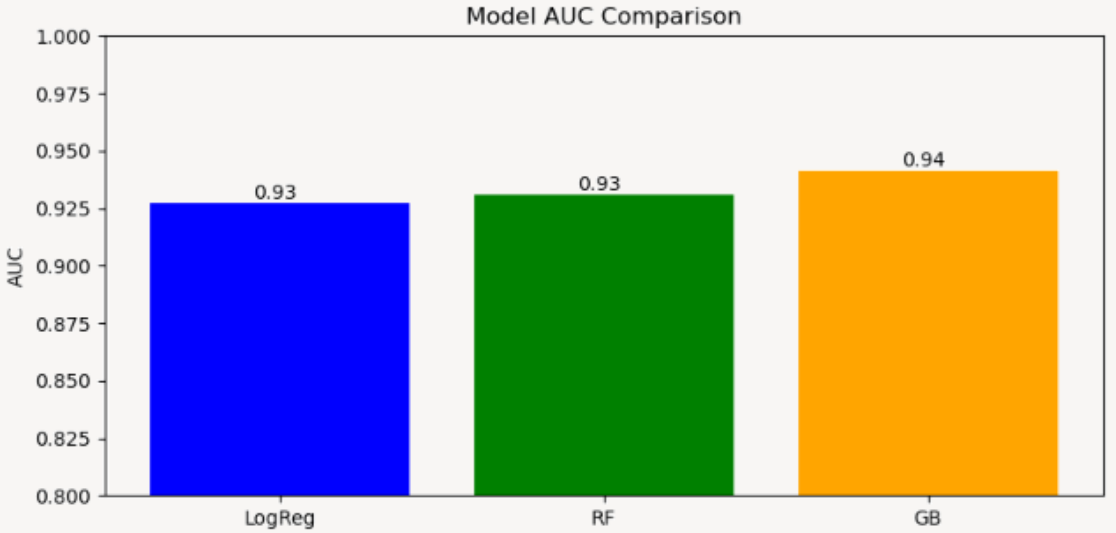
Solving For:

Account	Whether the individual has an account at a financial institution, a mobile money account or both (=1 or 0) **– Binary classification**

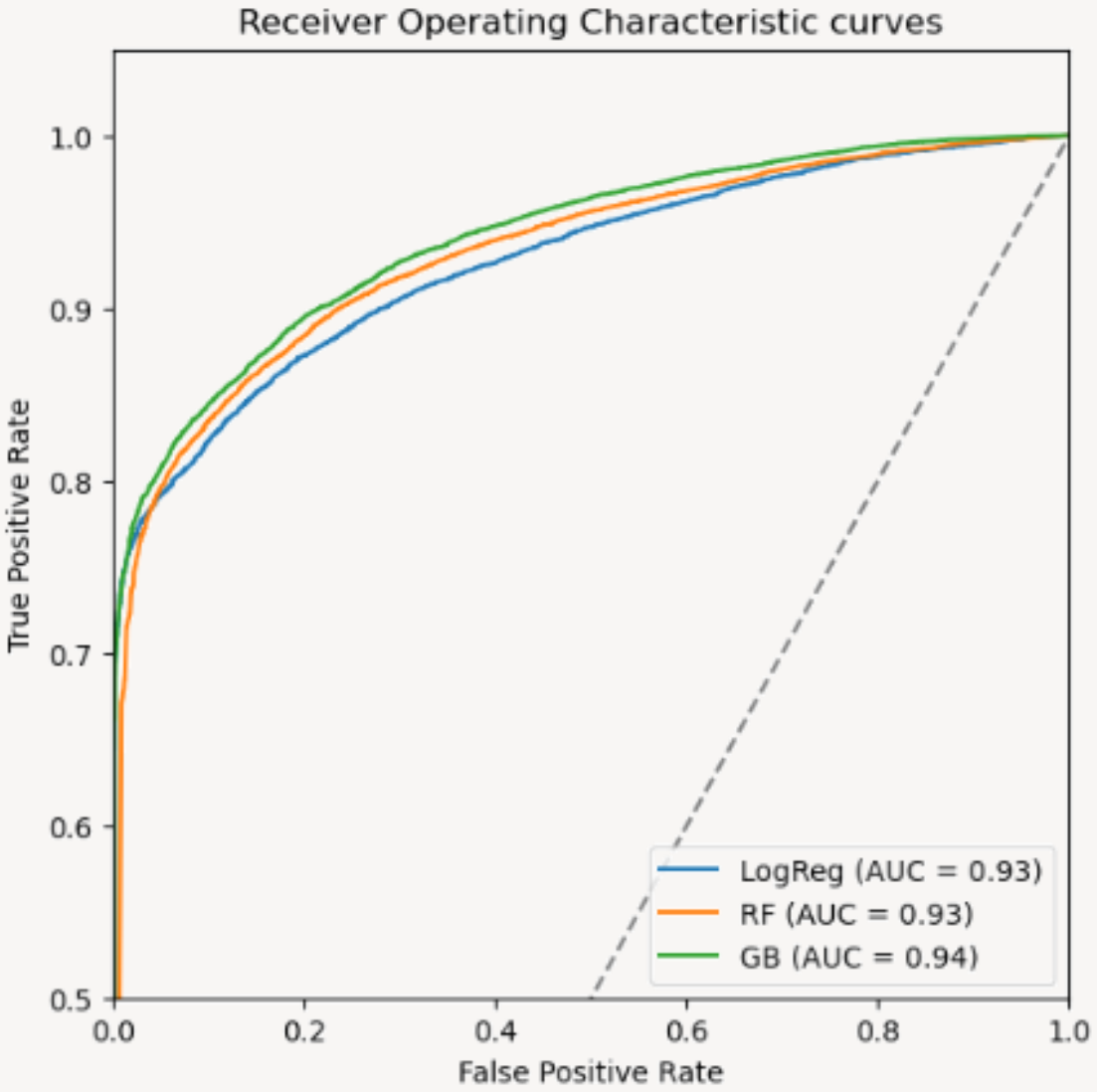
Comparative Evaluation of Classification Models

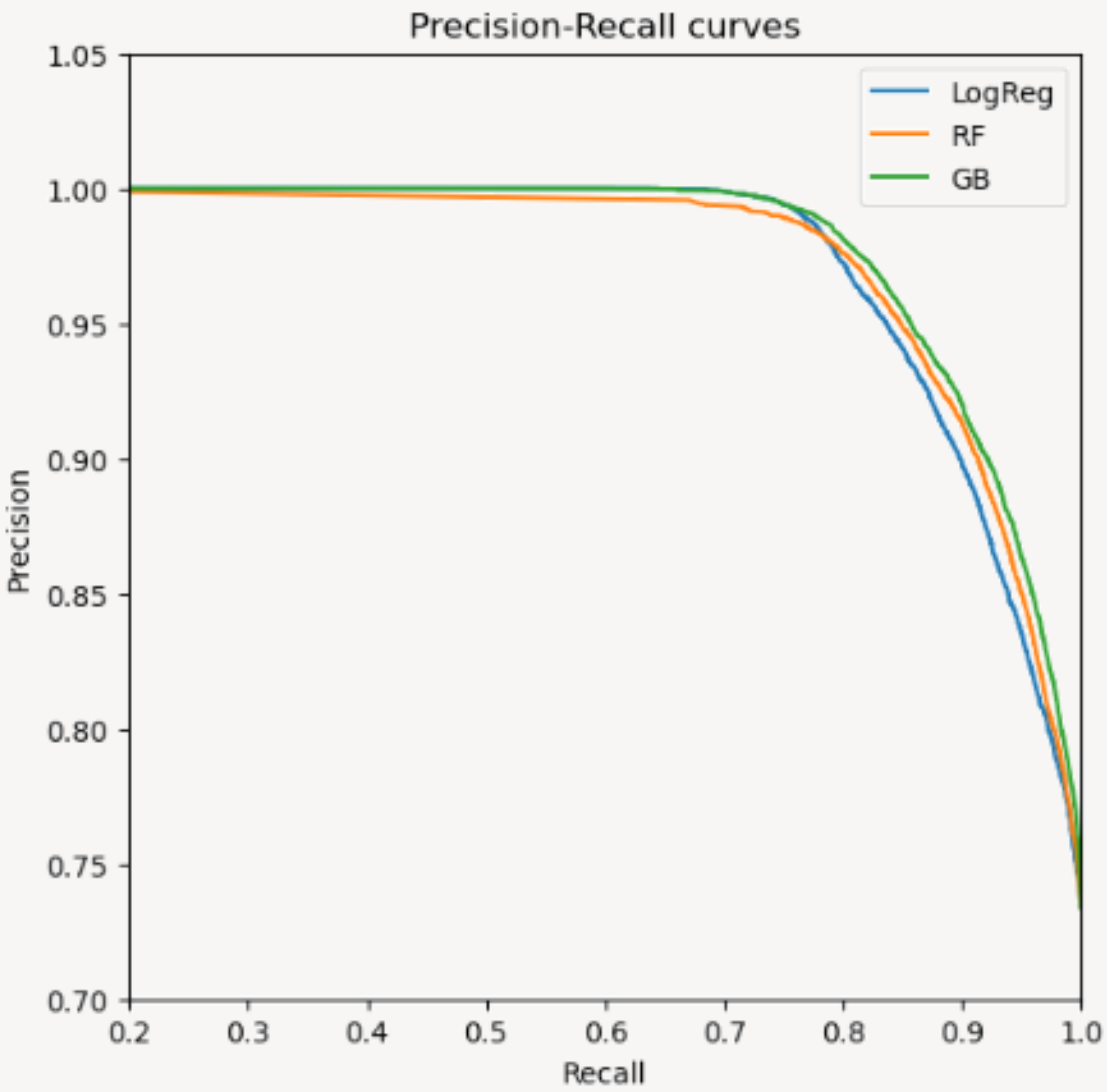
Model Type	Logistic Regression	Random Forest	Gradient Boosting
Accuracy:	0.85	0.86	0.87
Precision:	0.91	0.91	0.92
Recall:	0.88	0.90	0.90
F1 Score:	0.90	0.91	0.91
AUC:	0.93	0.93	0.94
Log Loss:	0.28	0.47	0.26





Comparative Evaluation of Classification Models





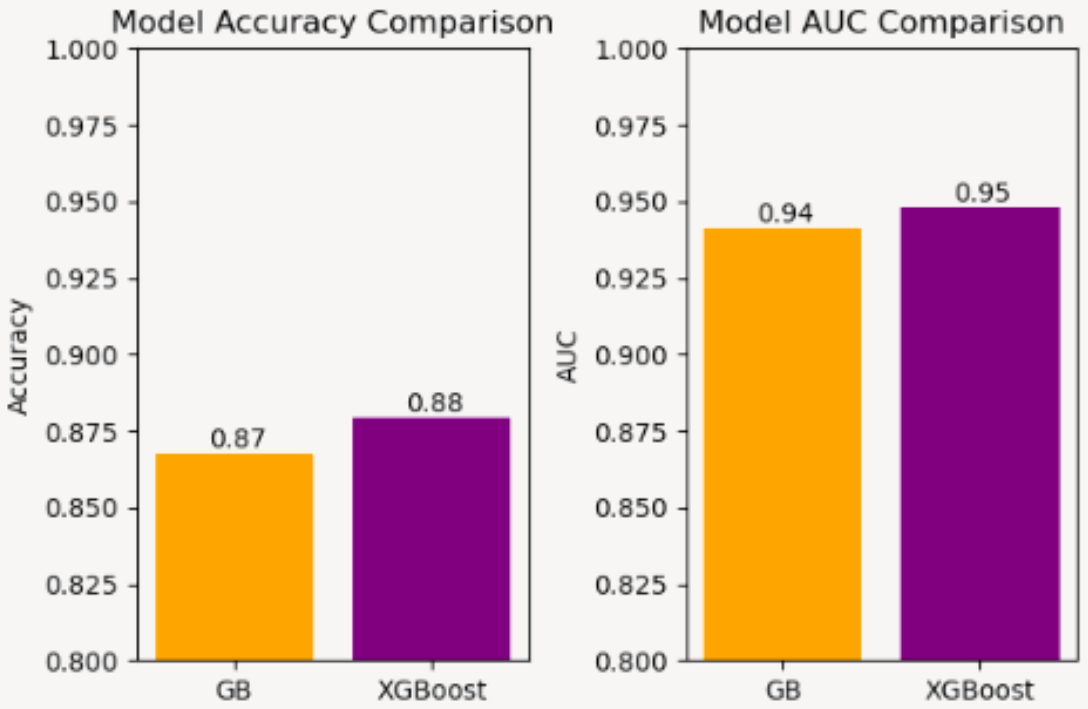
But The Real Winner Is…

Model Type	Gradient Boosting	XGBoost
Accuracy:	0.87	0.88
Precision:	0.92	0.93
Recall:	0.90	0.91
F1 Score:	0.91	0.92
AUC:	0.94	0.95
Log Loss:	0.26	0.24

What is XGBoost?

Short for eXtreme Gradient Boosting. XGB is a more advanced and higher performing version of gradient boosting:

- XGB handles sparse or missing data more effectively
- Includes additional regularization to prevent overfitting
- Includes built-in cross validation at each iteration
- Enables parallel processing making it faster than normal GB



Next Steps

Split Dataset into High Income and Other Countries

- Create a more granular view on how well the model performs across different cohorts
- Intuitively, it is easier for the model to perform on high income countries (which have more robust data) than low-income countries, where behaviors are harder to predict.

Further Hyperparameter Tuning

- Improve model performance, accuracy and reduce log loss
- Use GridSearchCV function to tune the hyperparameters to find the set that gives the best performance
- Furthermore, high income vs. low income cohorts might require different hyperparameters for the model to perform

Create a Financial Inclusion Score

- Create a standardized score using techniques such as linear scaling, sigmoid calibration on the predicted probabilities
- This may have real-world value by ascribing a metric to assess an individual's eligibility for credit and financial inclusion

Thank you

Yaksh Birla

yakshb@outlook.com

LinkedIn