Machine Learning Analysis on Contextual and Covariate Features in Satellite Image of Lagos, Nigeria

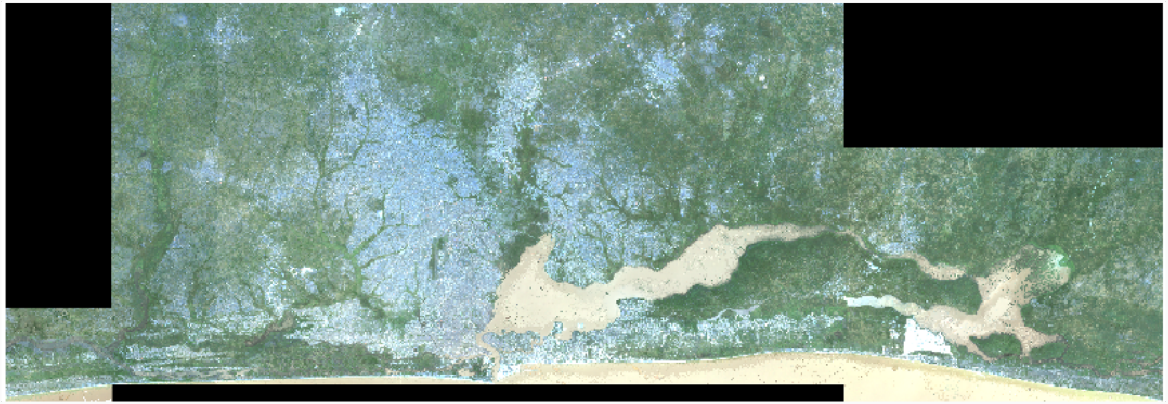
Data Science Capstone Spring 2022

By Jake Lieberfarb and Bradley Reardon

Supervisors: Professor Dr. Ryan Engstrom
Professor Dr. Amir Jafari

Table of Contents

Original Raw Image of Lagos, Nigeria

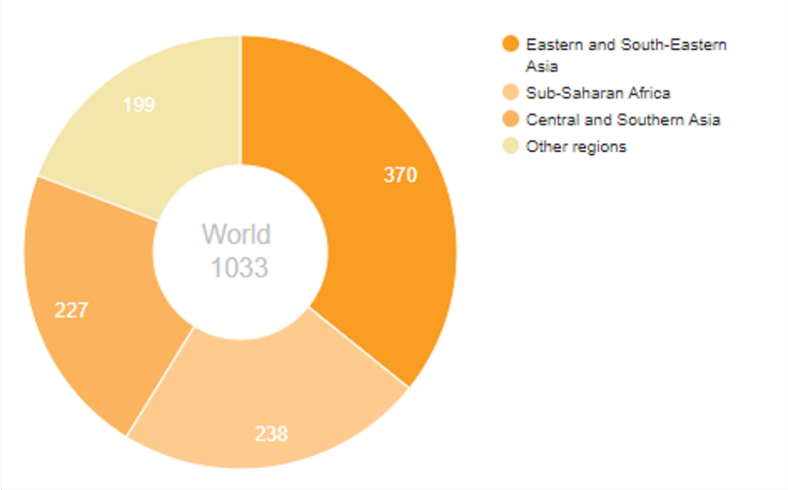


QR code to our github

Problem Statement

Background:

- According to the UN, more than 1 billion people are living in deprived areas
- Policy makers, government and global organizations are seeking detailed identification of deprived areas to enhance assignment of resources and track progress of development projects



There were two questions the researchers hoped to address in this study:

1. **Can feature importance methods derive any contextual or covariate features that are useful in identifying 'Deprive' and 'Built-up' areas?**

2. **Through the use of classical machine learning models and statistical analysis, are contextual and/or covariate features useful in identifying 'Deprived' and 'Built-up' areas?**

*https://unstats.un.org/sdgs/report/2019/goal-11/

The Data

Labeled Areas

- 47,560 labeled areas in Lagos
- Three labels: **Deprived, Built-up** and **Non-Built-up**
- 10 m^2 per label
- Highly imbalanced

Raw Satellite Image

- GeoTiff file for Lagos

Contextual Features

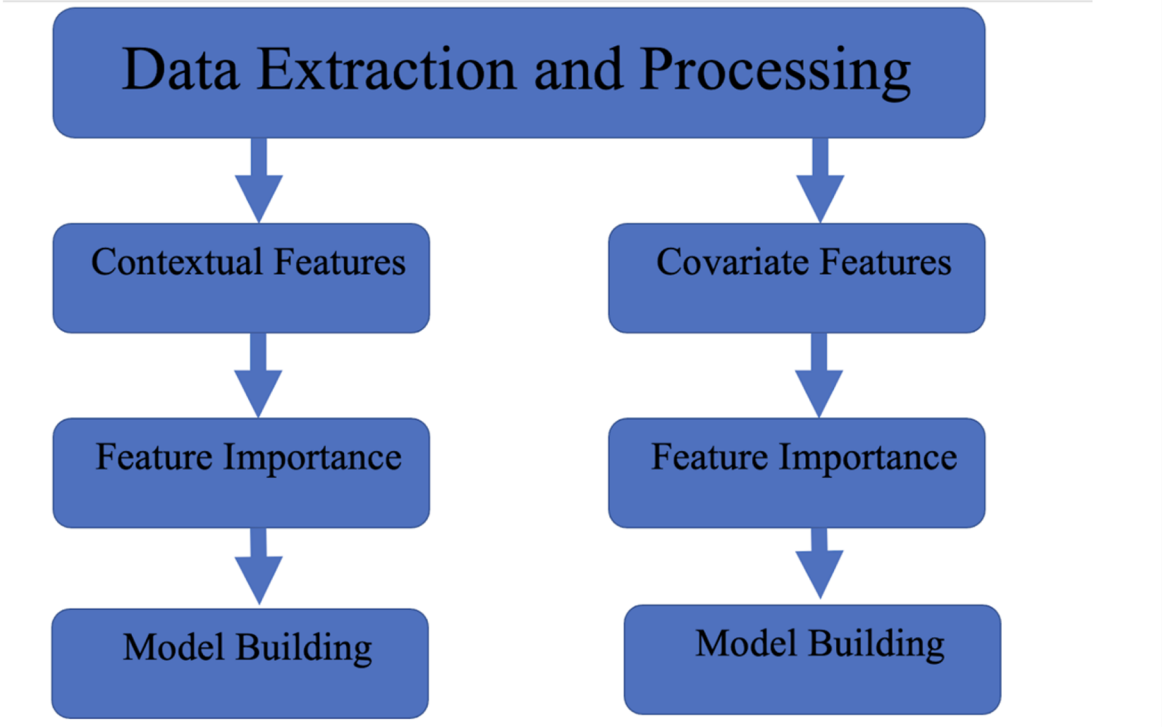
- 144 contextual features
- GeoTiff file per each feature

Covariate Features

- Pixel based
- Format: GeoTiff file

Flow of Analysis

- Contextual and Covariate features followed the same:
  - Preprocessing
  - Feature importance procedure
  - Model building method
  - Standardized and split on 60/20/20
  - Same hyperparameter features and cross validation
  - **Random state** = 42
  - **Alpha** = 0.05

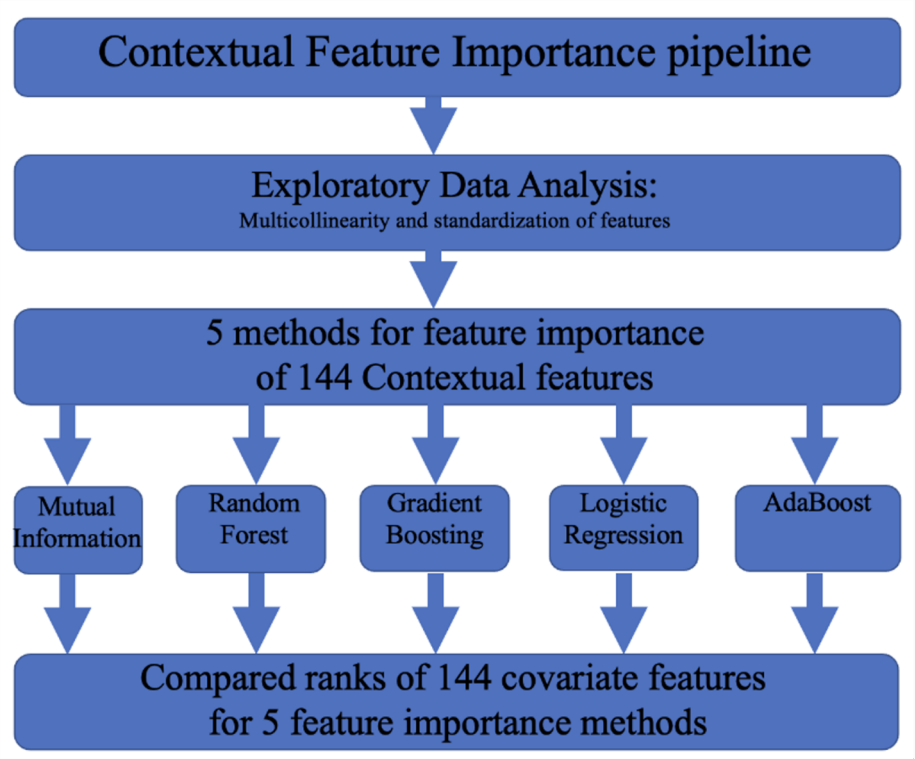


Contextual Features

Table of Contents

Original Raw Image of Lagos, Nigeria

Flow of Contextual Feature Importance Analysis

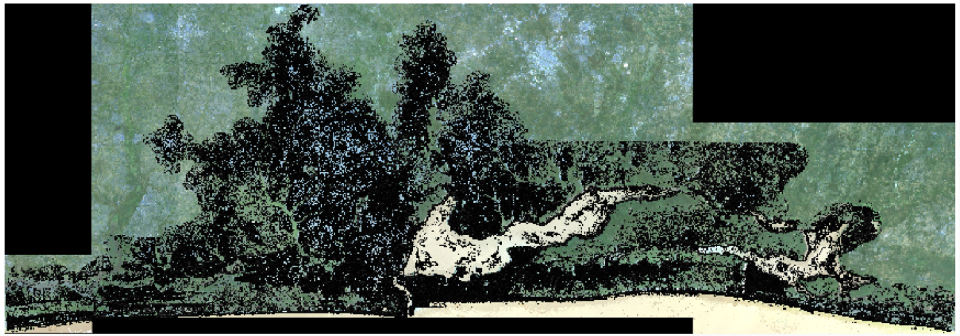


Extracting Contextual Features (Terminology)

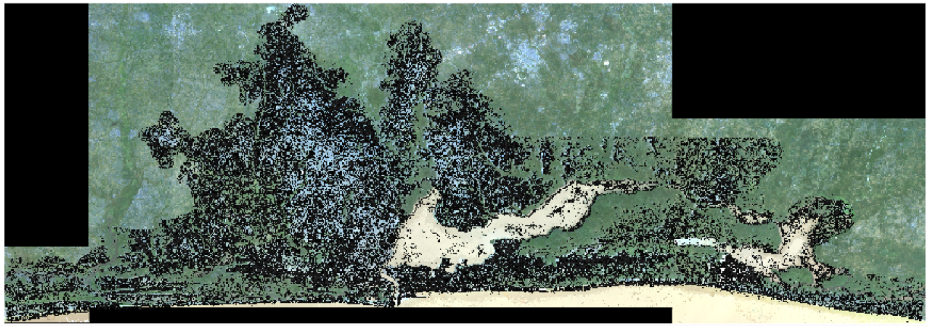
Contextual features-144 statistical measurements for each pixel taken from satellite imaging

- Contextual features were defined as the statistical quantification of edge patterns, pixel groups, gaps, textures, and the raw spectral signatures calculated over groups of pixels or neighborhoods.
- Examples:

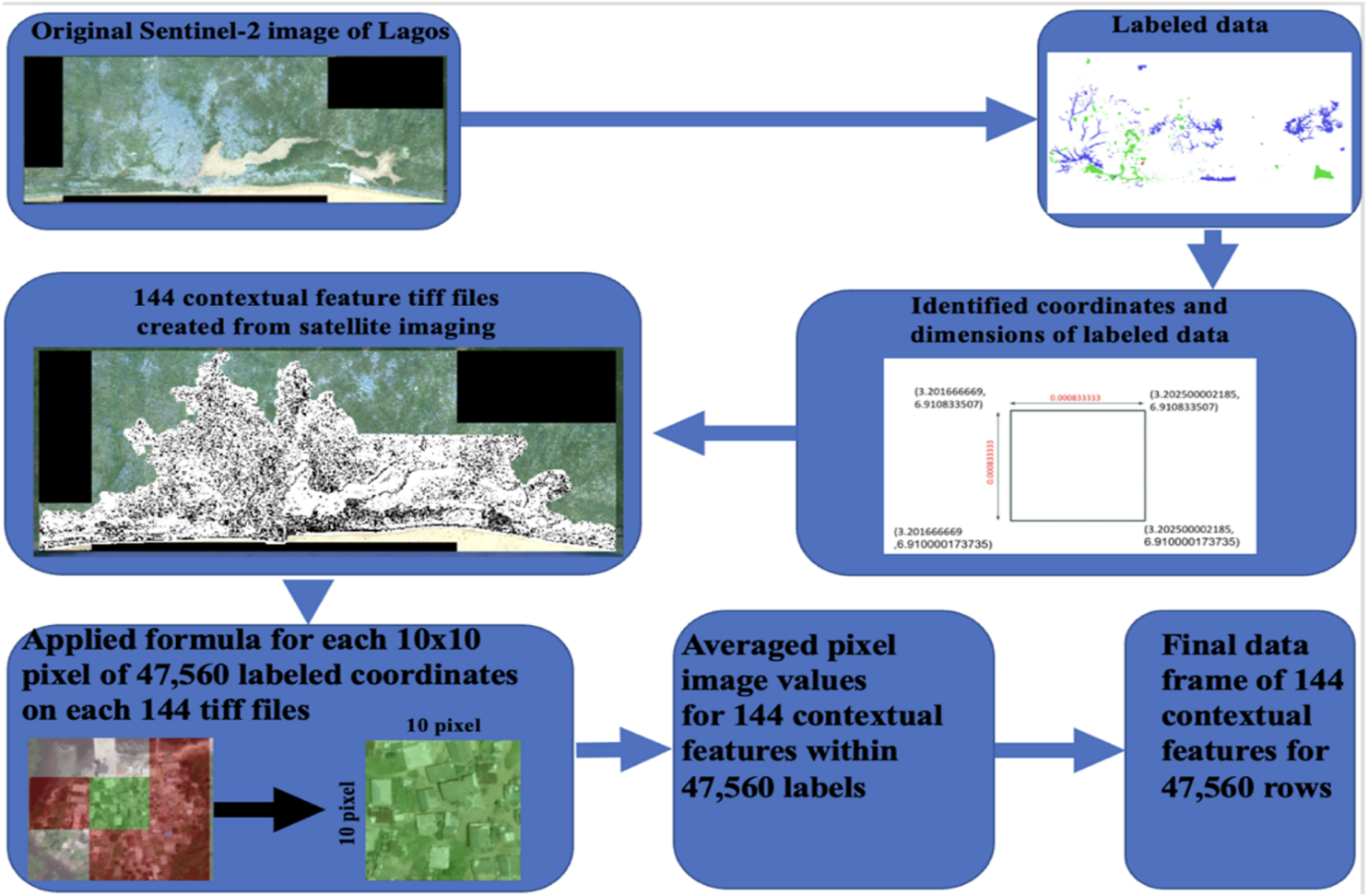
sfs_sc31_w_mean



pantex_sc7_min

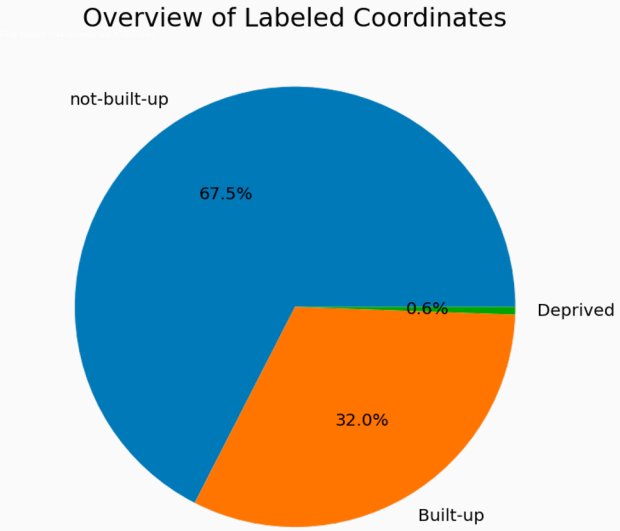


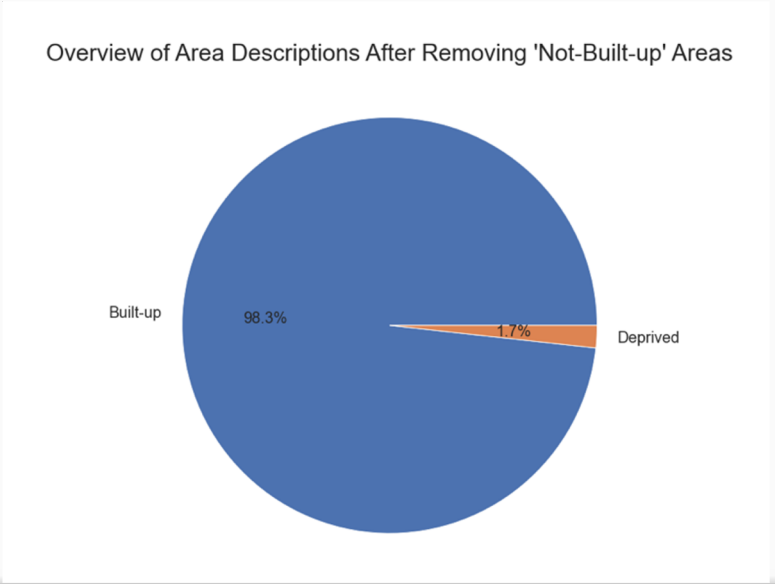
Extracting Contextua Features



Preprocessing the Contextual Features (2/2)

- Removed ‘Not-Built-up’ from the dataset
  - New dataframe:
    - 15471 samples
    - 98.3% - ‘Built-up’
    - 1.7% - ‘Deprived’

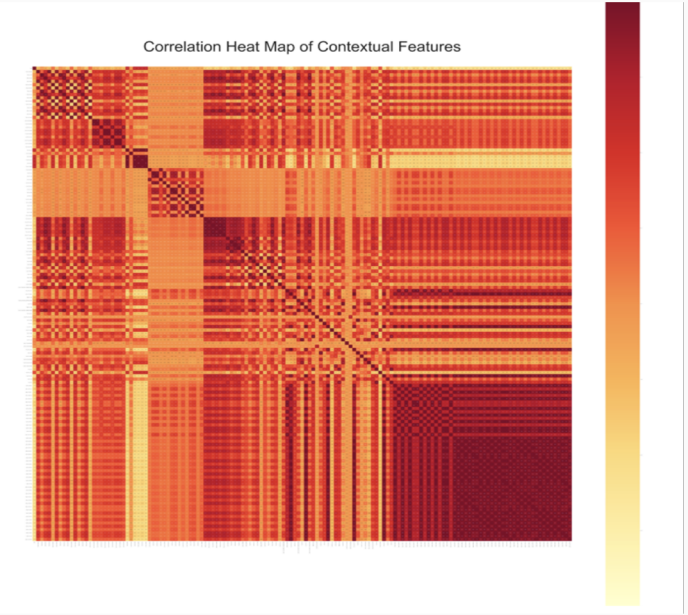


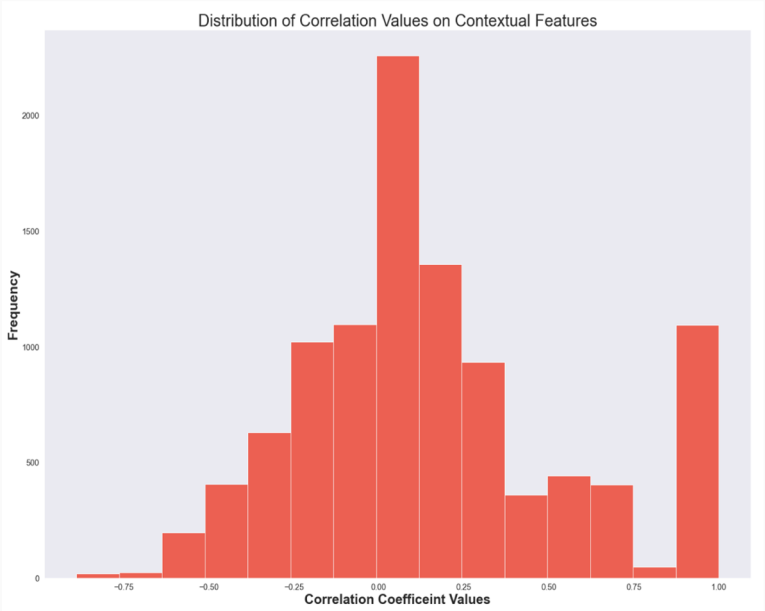


Contextual Features Correlation Plot

Many of the variables had a strong correlation with each other as indicated by **dark red** and **light-yellow** areas

- Histogram showed strong negative and positive correlation
- Many features had low correlation with each other (**skew = 0.579909**)



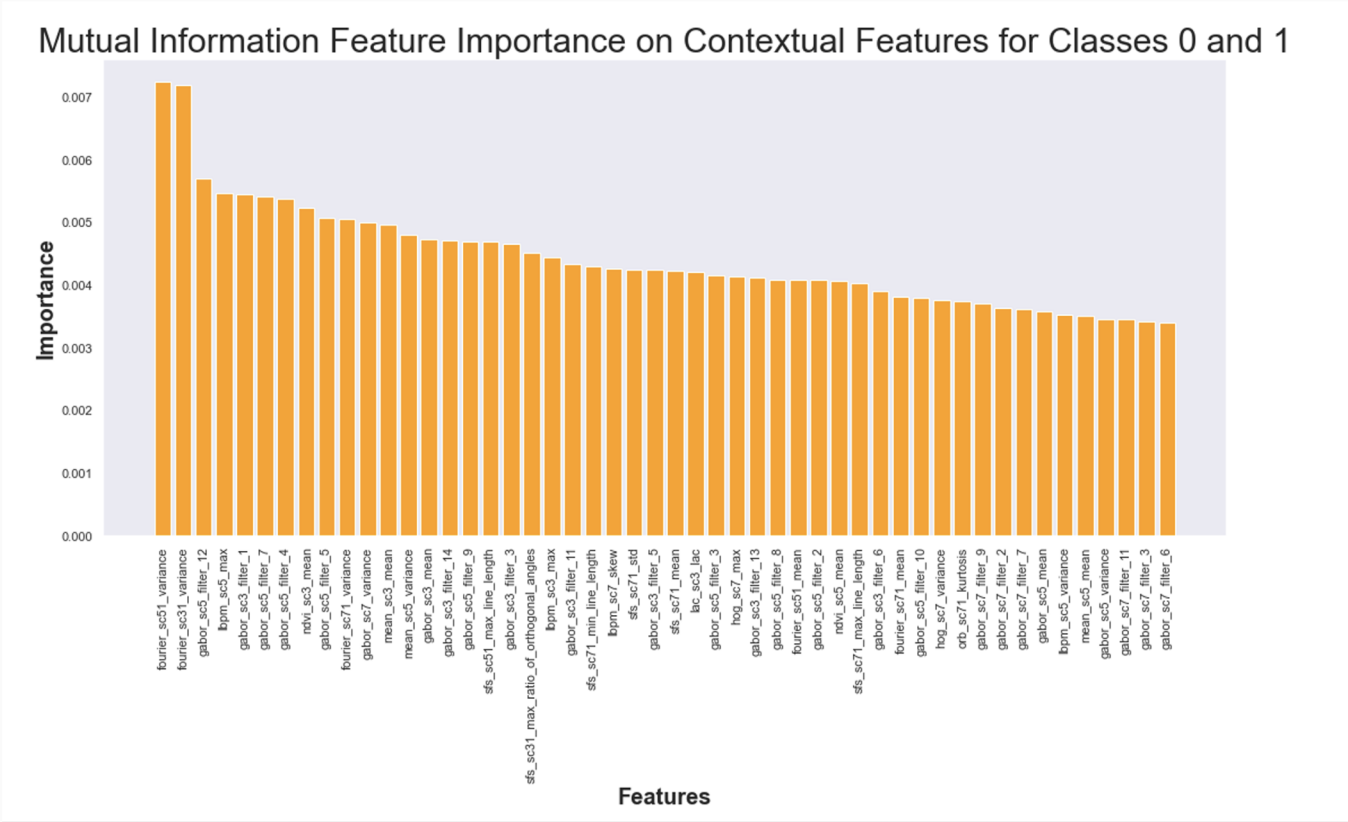


Feature Importance: Methodology

Feature Importance Method	***Sklearn Function***	***Methodology***
Mutual Information	*SelectKBest, mutual_info_classif*	*nonparametric methods related to entropy estimation from k-nearest neighbors distances. Mutual information is closely related to entropy and provides results from the range of zero to 1.*
Random Forest	*feature_importances_*	*The mean and standard deviation of accumulation of the impurity decrease within each decision tree of the random forest*
Logistic Regression	*LogisticRegression*	*odds ratio for coefficients*
Gradient Boosting	*feature_importances_*	*rank features based on the total reduction of the criterion within each feature (Gini importance)*
Adaboost	*feature_importances_*	*rank features based on the total reduction of the criterion within each feature (Gini importance)*

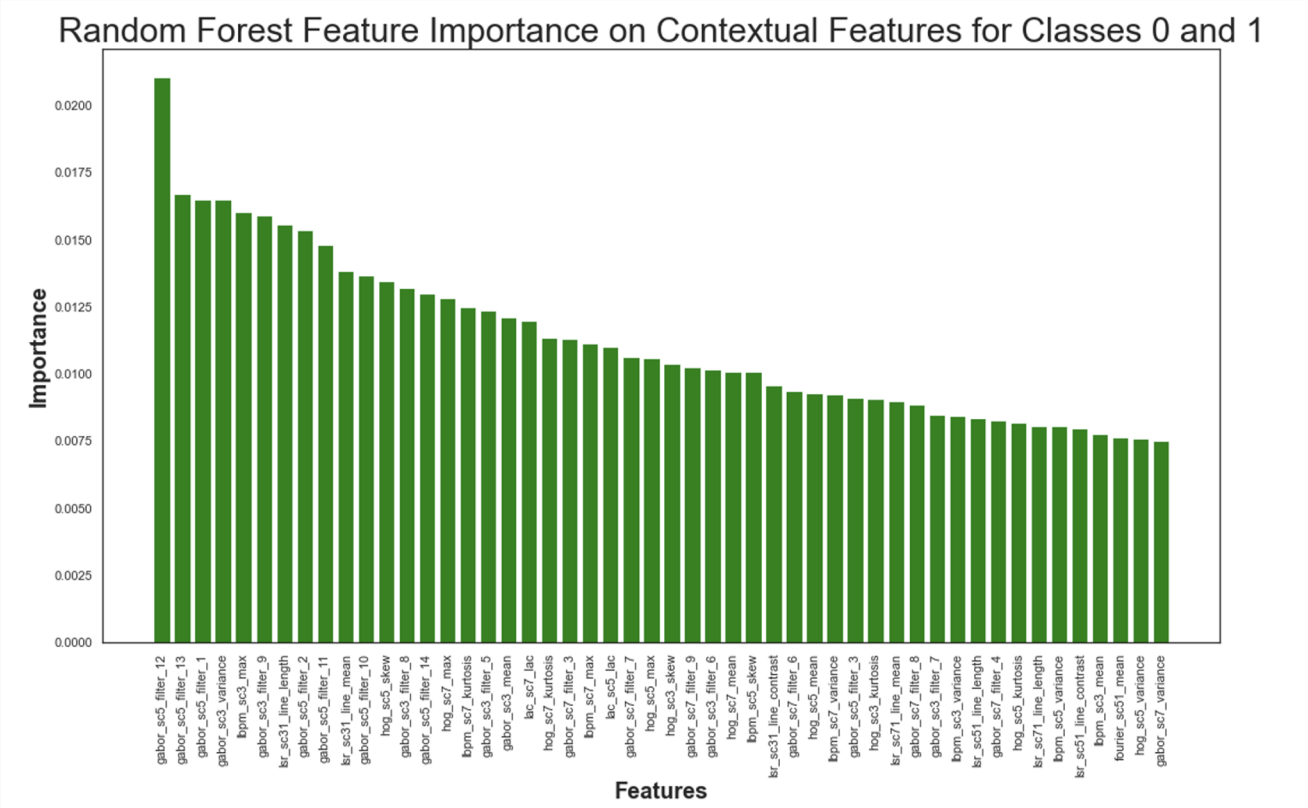
Feature Ranking: Mutual Information

- Feature importance conducted on validation set
- Most significant feature:
  - ‘fourier_sc51_variance’



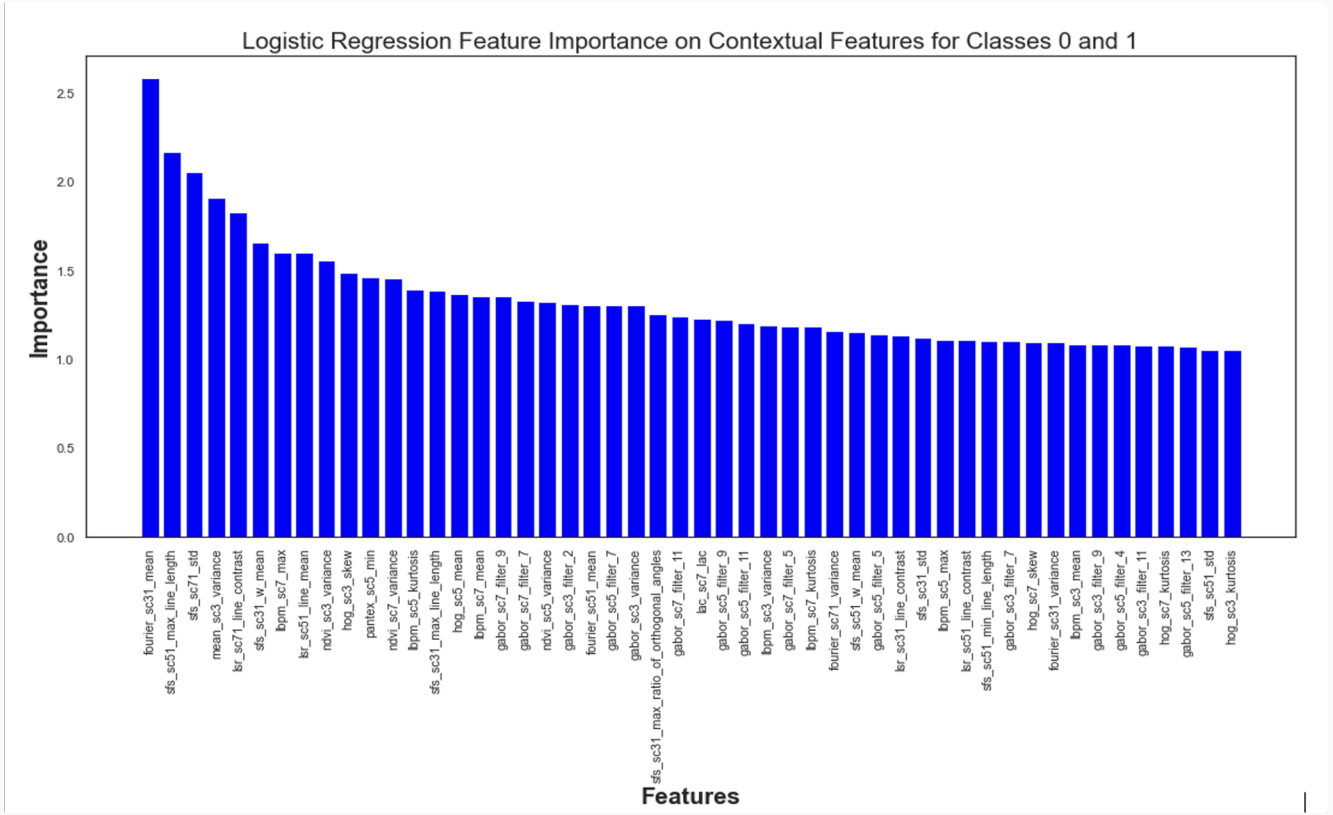
Feature Ranking: Random Forest

- Feature importance conducted on validation set
- Most significant feature: ‘gabor_sc5_filter_12’



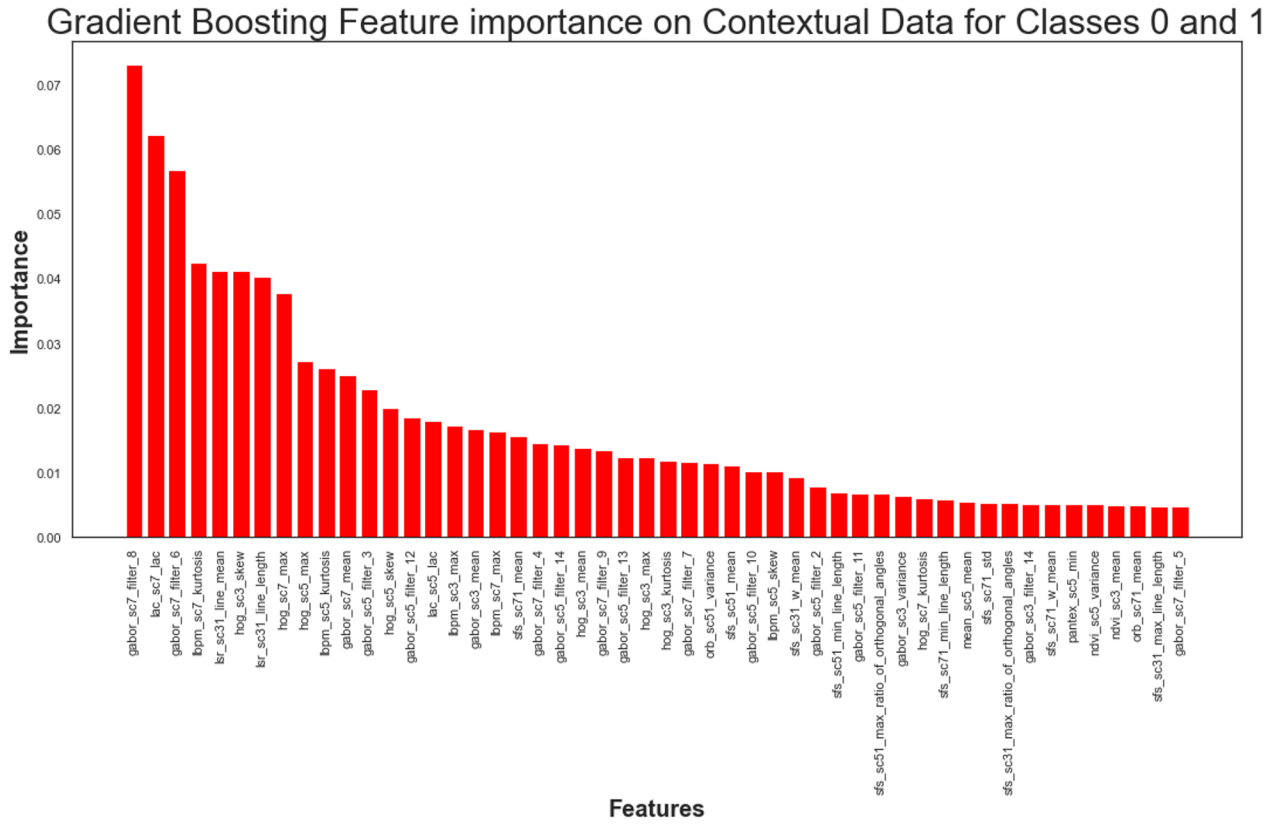
Feature Ranking: Logistic Regression

- Feature importance conducted on Validation set
- Most significant feature:
  - ‘fourier_sc31_mean’



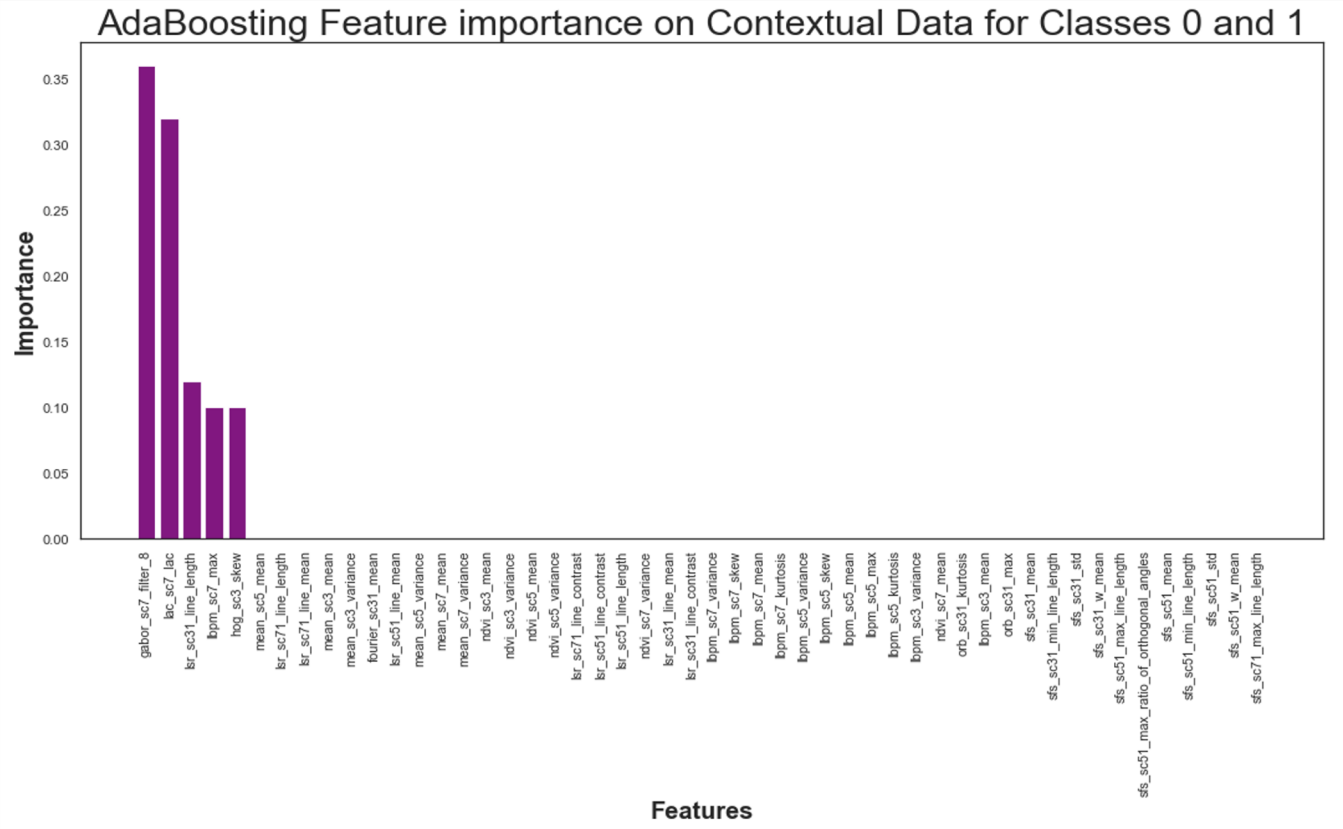
Feature Ranking: Gradient Boosting

- Feature importance conducted on Validation set
- Most significant feature:
  - ‘gabor_sc7_filter_8’



Feature Ranking: Adaptive Boosting

- Feature importance conducted on Validation set
- If feature was not significant, it was given value of zero
  - 5 features were important
- Most significant feature:
  - ‘gabor_sc7_filter_8



Feature Ranking: Results

- **No clear pattern was detected amongst Contextual Feature methods**
  - **Feature with rank 1: ‘lbpm_sc7_max’**

Model	Validation F1 for ‘Deprived’
Random Forest	0.07
Logistic Regression	0.07
Gradient Boosting	0.00
Adaboost	0.00

Feature Ranking: Comparing all Ranks

- Contextual features would have feature importance values for some models but have low values for others

Contextual_ features	top_logistic _0_1	top_ Random _Forest_0_1	top_ Gradient_ Boosting_0_1	top_ Ada _Boosting_0_1	minfo_0_1	rank
lbpm_sc7_max	7	22	18	4	39	1
hog_sc3_skew	10	26	6	5	73	2
gabor_sc5_filter_12	69	1	14	82	3	3
lac_sc7_lac	26	19	2	2	131	4
gabor_sc7_filter_8	91	38	1	1	53	5

Model Development: Contextual - classes 0 & 1

- 5 model types tested:
  - Ensemble (9 MLP models)
  - Multilayer Perceptron
  - Random Forest
  - LogisticRegression
  - Gradient Boosting
- train/test/val split: 60/20/20
- GridsearchCV used for hyperparameter tuning
- Top 50 features in each feature reduction set tested
- Best performing model:
  - Ensemble model
  - Using all features available

Dataset	Model	Feature Set	Feature Count	Classes	F1 - Class 1 (Deprived)	F1 - Macro
contextual	Ensemble	All_Features	144	classes_0&1	0.35	0.67
contextual	Gradient_Boosting	ADA_Features	50	classes_0&1	0.09	0.54
contextual	Gradient_Boosting	All_Features	144	classes_0&1	0.09	0.54
contextual	Random_Forest	ADA_Features	50	classes_0&1	0.07	0.53
contextual	Gradient_Boosting	Gradient_Boosting_Features	50	classes_0&1	0.07	0.53
contextual	MLP	Gradient_Boosting_Features	50	classes_0&1	0.04	0.51
contextual	MLP	All_Features	144	classes_0&1	0.04	0.51
contextual	MLP	ADA_Features	50	classes_0&1	0.04	0.51
contextual	MLP	Logistic_Features	50	classes_0&1	0.04	0.51
contextual	Logistic_Regression	All_Features	144	classes_0&1	0.04	0.51
contextual	Logistic_Regression	ADA_Features	50	classes_0&1	0.04	0.51
contextual	Logistic_Regression	Logistic_Features	50	classes_0&1	0.04	0.51
contextual	Random_Forest	All_Features	144	classes_0&1	0.04	0.51
contextual	Random_Forest	Logistic_Features	50	classes_0&1	0.04	0.51
contextual	Logistic_Regression	Gradient_Boosting_Features	50	classes_0&1	0.04	0.08
contextual	Gradient_Boosting	Minfo_Features	50	classes_0&1	0.03	0.51
contextual	Gradient_Boosting	Logistic_Features	50	classes_0&1	0.03	0.51
contextual	MLP	Random_Forest_Features	50	classes_0&1	0.03	0.50
contextual	MLP	Minfo_Features	50	classes_0&1	0.00	0.50
contextual	Gradient_Boosting	Random_Forest_Features	50	classes_0&1	0.00	0.49
contextual	Logistic_Regression	Random_Forest_Features	50	classes_0&1	0.00	0.50
contextual	Logistic_Regression	Minfo_Features	50	classes_0&1	0.00	0.50
contextual	Random_Forest	Random_Forest_Features	50	classes_0&1	0.00	0.50
contextual	Random_Forest	Gradient_Boosting_Features	50	classes_0&1	0.00	0.50
contextual	Random_Forest	Minfo_Features	50	classes_0&1	0.00	0.50

Best Model: Contextual - classes 0 & 1

BEST PERFORMING CONTEXTUAL MODEL ON ALL CLASSES

Dataset	*Model*	*Feature Set*	*Feature Count*	*Classes*	*F1-Deprived*	*F1-Macro*
Contextual	*Ensemble*	*All*	*144*	*classes 0&1*	*0.35*	*0.67*

a. Ensemble model consisted of nine (9) MLP models

CONTEXTUAL CONFUSION MATRIX - CLASSES 0 & 1

True	***Built-up***	*3033*	*9*
	Deprived	*39*	*13*
		Built-up	***Deprived***
	Predicted		

(Confusion Matrix for the best performing model trained on the contextual features dataset containing only instances of the built-up (0) and deprived (1) classes.)

Model Parameters

- Hidden layer sizes: (100, 100, 100)
  - Number of neurons in hidden layer
- Activation: tanh
  - the hyperbolic tan function, returns f(x) = tanh(x)
- Solver: adam
  - refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba
- Alpha: 0.0001
  - L2 penalty (regularization term) parameter
- Learning Rate: invscaling
  - gradually decreases the learning rate at each time step ‘t’ using an inverse scaling exponent of ‘power_t’. effective_learning_rate = learning_rate_init / pow(t, power_t)

Model Testing: Contextual - all classes

- 5 model types tested:
  - Ensemble (9 MLP models)
  - Multilayer Perceptron
  - Random Forest
  - LogisticRegression
  - Gradient Boosting
- train/test/val split: 60/20/20
- GridsearchCV used for hyperparameter tuning
- Top 50 features in each feature reduction set tested
- Best performing model:
  - Ensemble model
  - Using all features available

Dataset	Model	Feature Set	Feature Count	Classes	F1 - Class 1 (Deprived)	F1 - Macro
contextual	Ensemble	All_Features	144	all_classes	0.25	0.70
contextual	MLP	All_Features	144	all_classes	0.21	0.67
contextual	Gradient_Boosting	Minfo_Features	50	all_classes	0.18	0.67
contextual	MLP	ADA_Features	50	all_classes	0.10	0.64
contextual	MLP	Logistic_Features	50	all_classes	0.09	0.63
contextual	Random_Forest	ADA_Features	50	all_classes	0.08	0.63
contextual	Gradient_Boosting	ADA_Features	50	all_classes	0.07	0.62
contextual	Gradient_Boosting	All_Features	144	all_classes	0.07	0.63
contextual	Gradient_Boosting	Gradient_Boosting_Features	50	all_classes	0.06	0.61
contextual	Random_Forest	All_Features	144	all_classes	0.04	0.62
contextual	Gradient_Boosting	Logistic_Features	50	all_classes	0.03	0.61
contextual	Logistic_Regression	Random_Forest_Features	50	all_classes	0.00	0.33
contextual	Logistic_Regression	Minfo_Features	50	all_classes	0.00	0.10
contextual	Logistic_Regression	Gradient_Boosting_Features	50	all_classes	0.00	0.32
contextual	MLP	Random_Forest_Features	50	all_classes	0.00	0.38
contextual	MLP	Gradient_Boosting_Features	50	all_classes	0.00	0.44
contextual	MLP	Minfo_Features	50	all_classes	0.00	0.38
contextual	Gradient_Boosting	Random_Forest_Features	50	all_classes	0.00	0.57
contextual	Logistic_Regression	All_Features	144	all_classes	0.00	0.60
contextual	Logistic_Regression	ADA_Features	50	all_classes	0.00	0.59
contextual	Logistic_Regression	Logistic_Features	50	all_classes	0.00	0.59
contextual	Random_Forest	Random_Forest_Features	50	all_classes	0.00	0.44
contextual	Random_Forest	Gradient_Boosting_Features	50	all_classes	0.00	0.42
contextual	Random_Forest	Logistic_Features	50	all_classes	0.00	0.60
contextual	Random_Forest	Minfo_Features	50	all_classes	0.00	0.55

Best Model: Contextual - all classes

BEST PERFORMING CONTEXTUAL MODEL ON ALL CLASSES

Dataset	*Model*	*Feature Set*	*Feature Count*	*Classes*	*F1-Deprived*	*F1-Macro*
Contextual	*Ensemble*	*All*	*144*	*all classes*	*0.25*	*0.70*

a. Ensemble model consisted of nine (9) MLP models

CONTEXTUAL CONFUSION MATRIX - ALL CLASSES

True	***Built-up***	*2729*	*6*	*325*
	Deprived	*25*	*8*	*16*
	Non-built-up	*267*	*1*	*6138*
		Built-up	***Deprived***	***Non-built-up***
	Predicted			

(Confusion Matrix for the best performing model trained on the full contextual features dataset containing all instances of the Built-up (0), Deprived (1), and Non-built-up (2) classes)

Model Parameters

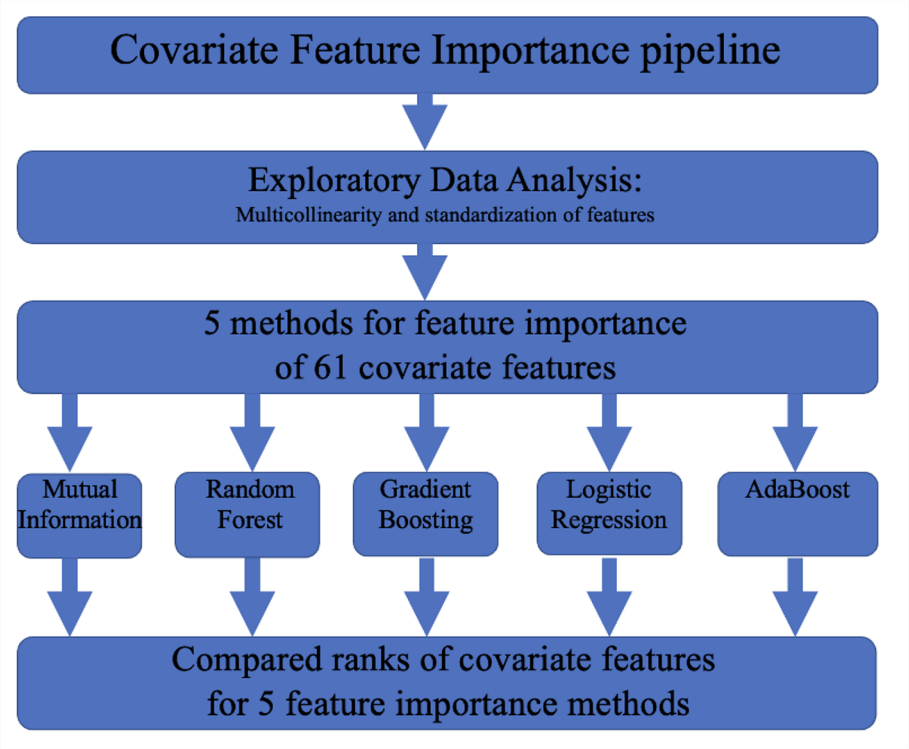
- Hidden layer sizes: (100, 100, 100)
  - Number of neurons in hidden layer
- Activation: tanh
  - the hyperbolic tan function, returns f(x) = tanh(x)
- Solver: adam
  - refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba
- Alpha: 0.0001
  - L2 penalty (regularization term) parameter
- Learning Rate: invscaling
  - gradually decreases the learning rate at each time step ‘t’ using an inverse scaling exponent of ‘power_t’. effective_learning_rate = learning_rate_init / pow(t, power_t)

Covariate Features

Table of Contents

Original Raw Image of Lagos, Nigeria

Flow of Covariate Feature Importance Analysis

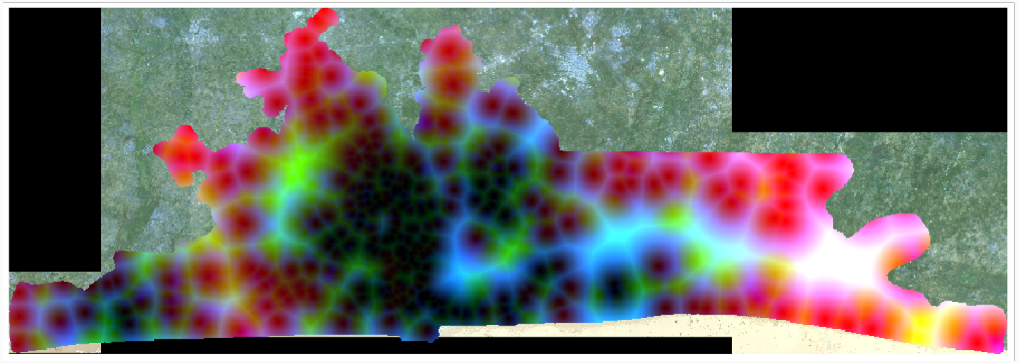


Extracting Covariate Features (Terminology)

Covariate features- 61 total:

- Features stored in a single geoTiff file
- 55 continuous variables
- 6 categorical variables

Covariate Features

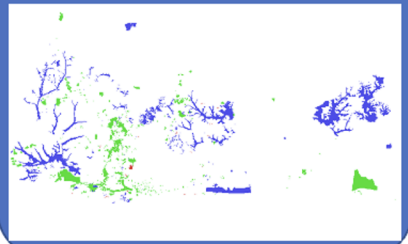


Covariate Feature	Count
Facilities & Services	8
Housing(HO)	1
Infrastructure	5
Physical Hazards & Assets	23
Population Counts	2
SES(HH)	14
Social Hazards and Assets	5
Unplanned Urbanization	3

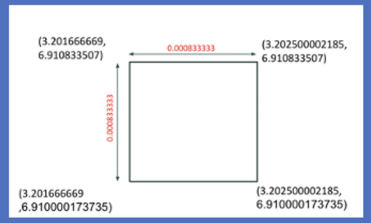
Extracting Covariate Features

Original Sentinel-2 image of Lagos

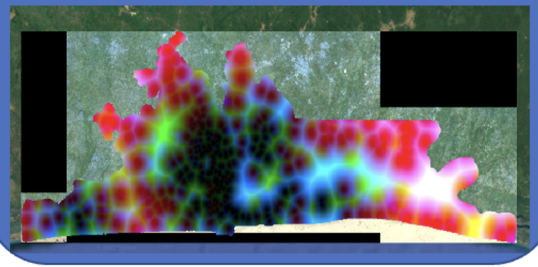
Labeled data



Identified coordinates and dimensions of labeled data



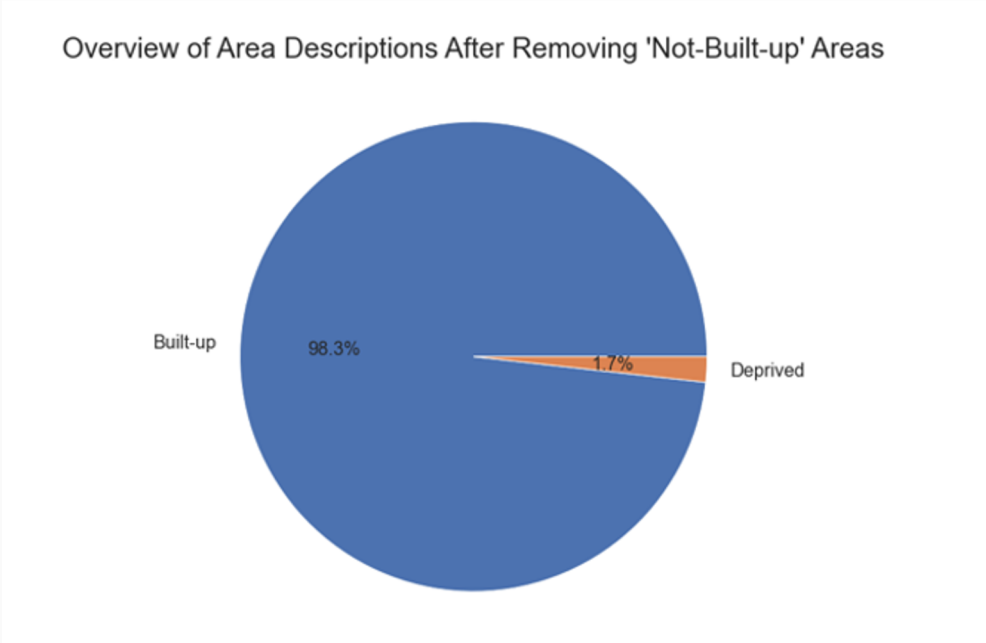
61 covariate features were labeled by distinct colors



Matched the covariate data to the 47,560 labeled data for each of the 61 covariate features to get final data frame

Preprocessing the Covariate Features

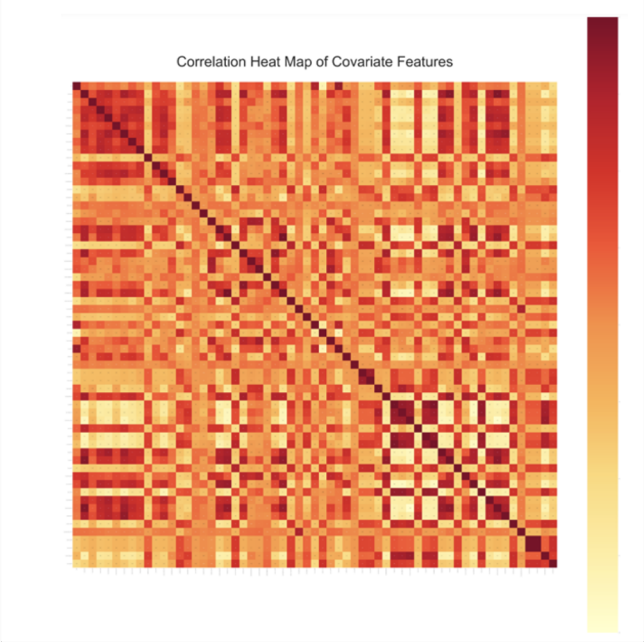
- Removed 1987 nan values
  - 1876 from ‘Not-Built-up’
  - 1 from ‘Built-up’
  - Removed ‘ph_gdmhz_2005’
    - Contained nan values
- Removed ‘Not-Built-up’ from the dataset
  - **New dataframe:**
    - 15470 samples
    - 98.3% - ‘Built-up’
    - 1.7% - ‘Deprived’

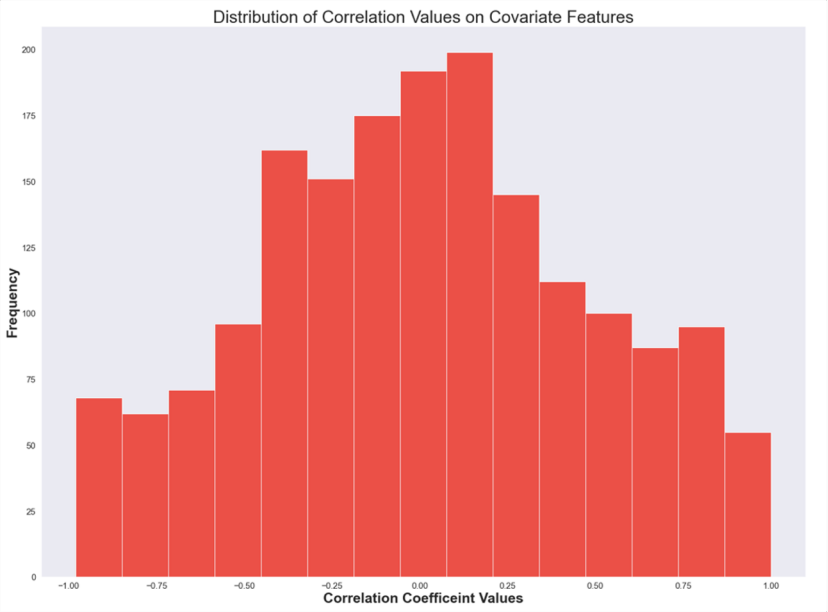


Covariate Features Correlation Plot

Many of the variables had a strong correlation with each other as indicated by **dark red** and **light-yellow** areas

- Histogram showed strong negative and positive correlation
- Many features had low correlation with each other (**skew = 0.017183**)



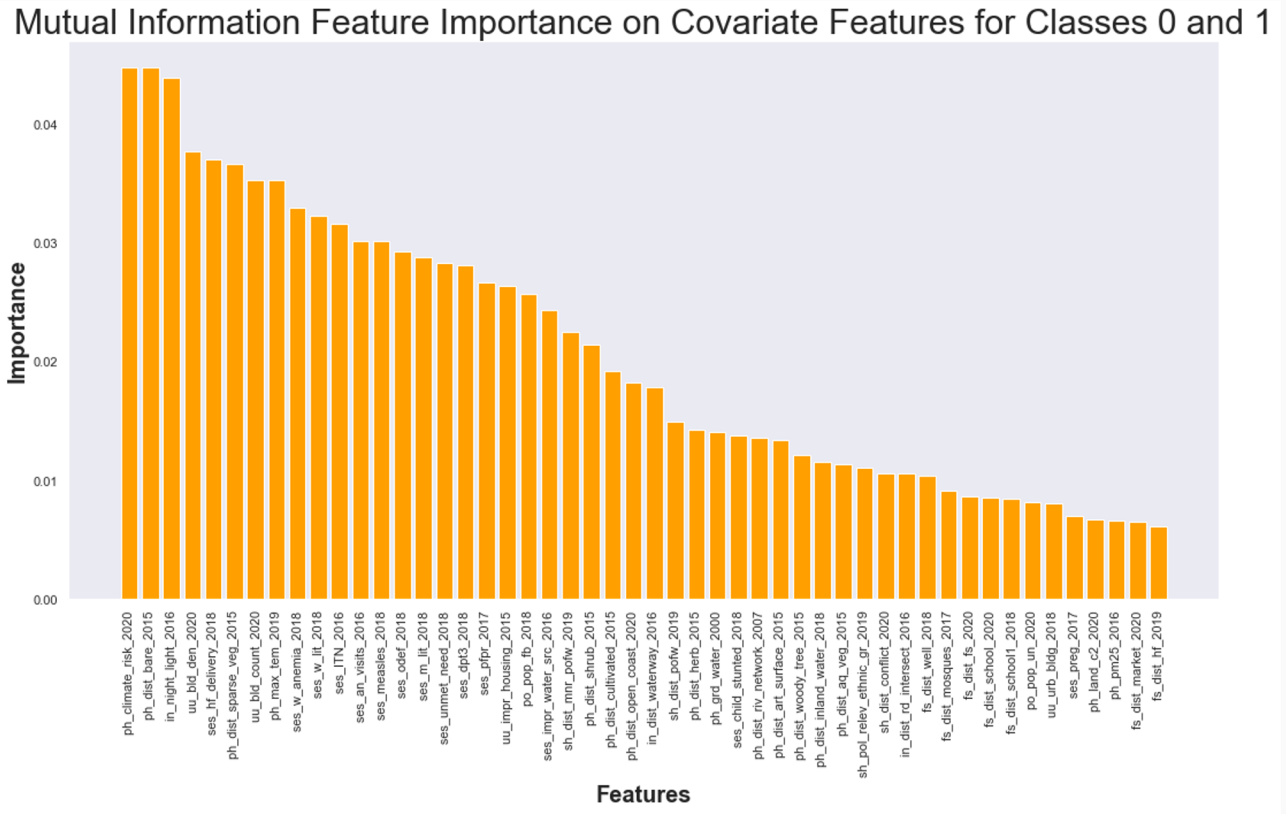


Feature Importance: Methodology

Feature Importance Method	***Sklearn Function***	***Methodology***
Mutual Information	*SelectKBest, mutual_info_classif*	*nonparametric methods related to entropy estimation from k-nearest neighbors distances. Mutual information is closely related to entropy and provides results from the range of zero to 1.*
Random Forest	*feature_importances_*	*The mean and standard deviation of accumulation of the impurity decrease within each decision tree of the random forest*
Logistic Regression	*LogisticRegression*	*odds ratio for coefficients*
Gradient Boosting	*feature_importances_*	*rank features based on the total reduction of the criterion within each feature (Gini importance)*
Adaboost	*feature_importances_*	*rank features based on the total reduction of the criterion within each feature (Gini importance)*

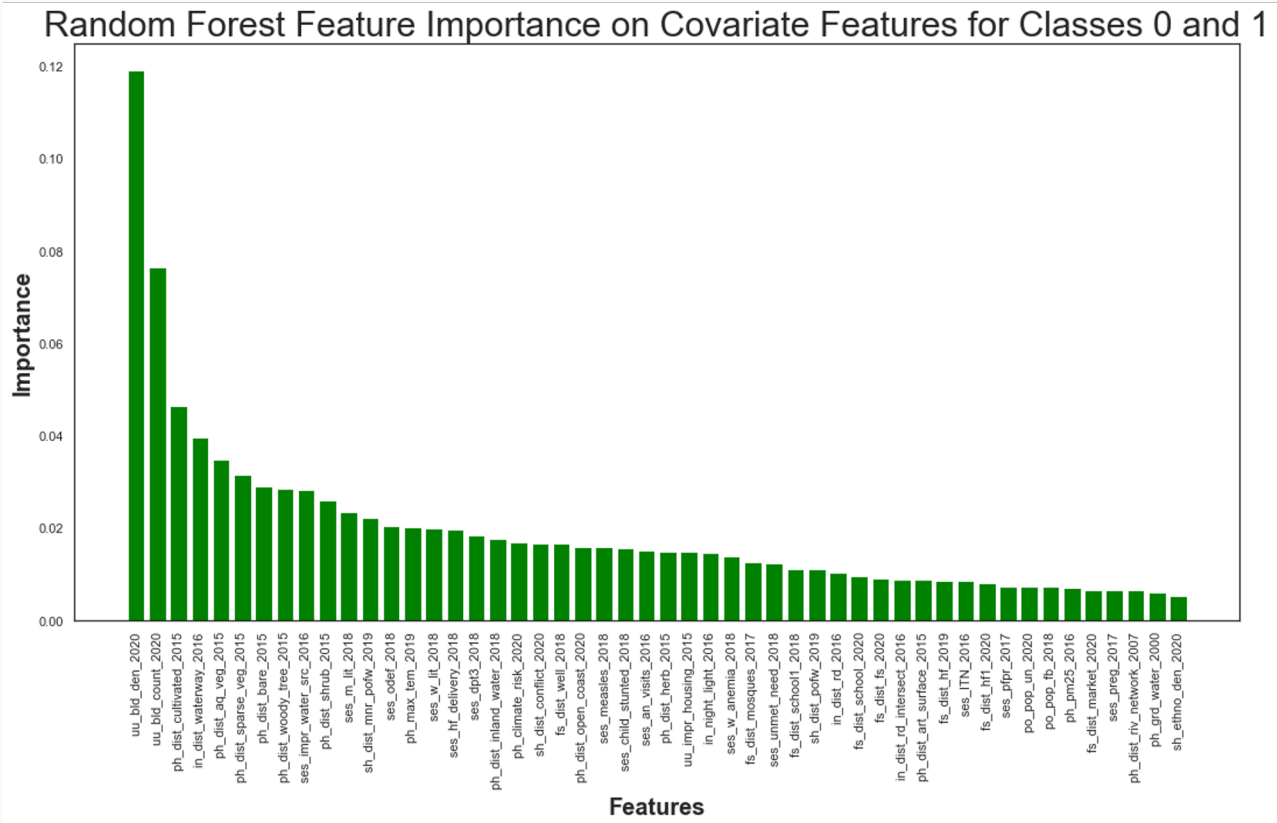
Feature Ranking: Mutual Information

- Feature importance conducted on validation set
- Most significant feature:
  - ‘ph_climate_risk_2020’



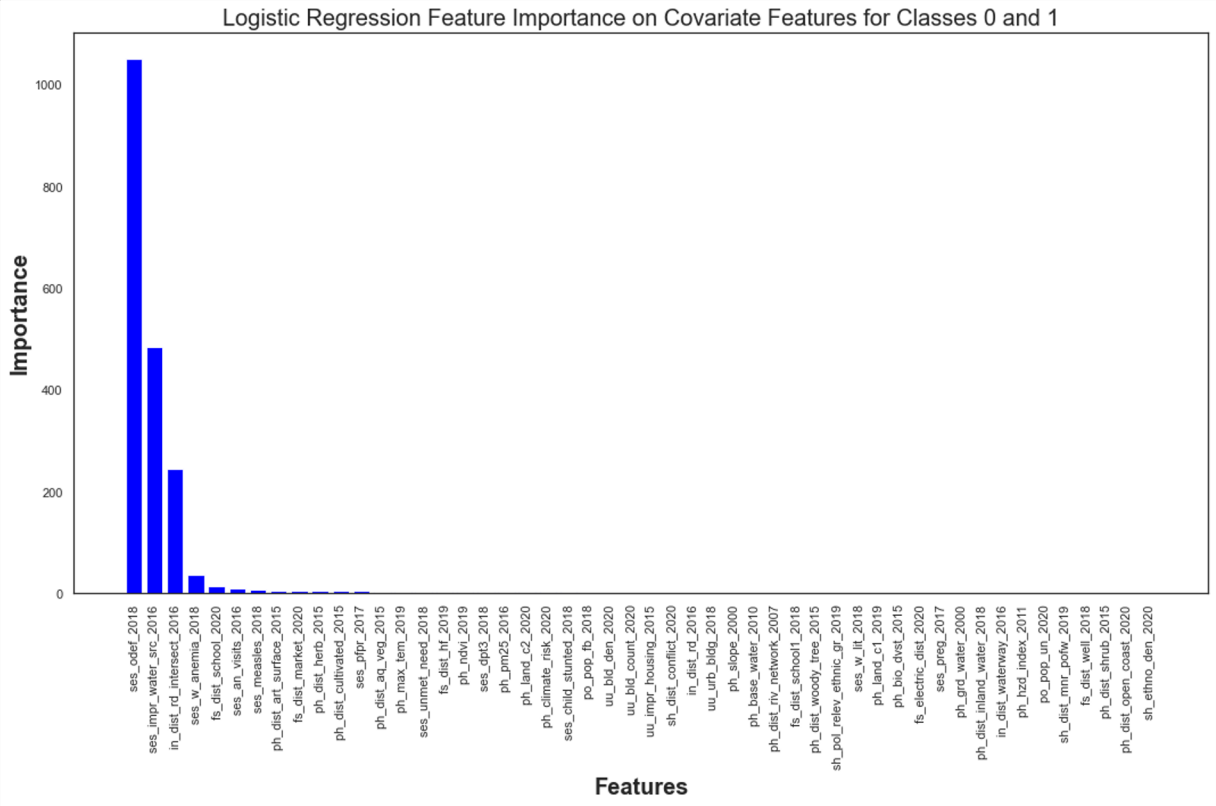
Feature Ranking: Random Forest

- Feature importance conducted on validation set
- Most significant feature: ‘uu_bld_den_2020’



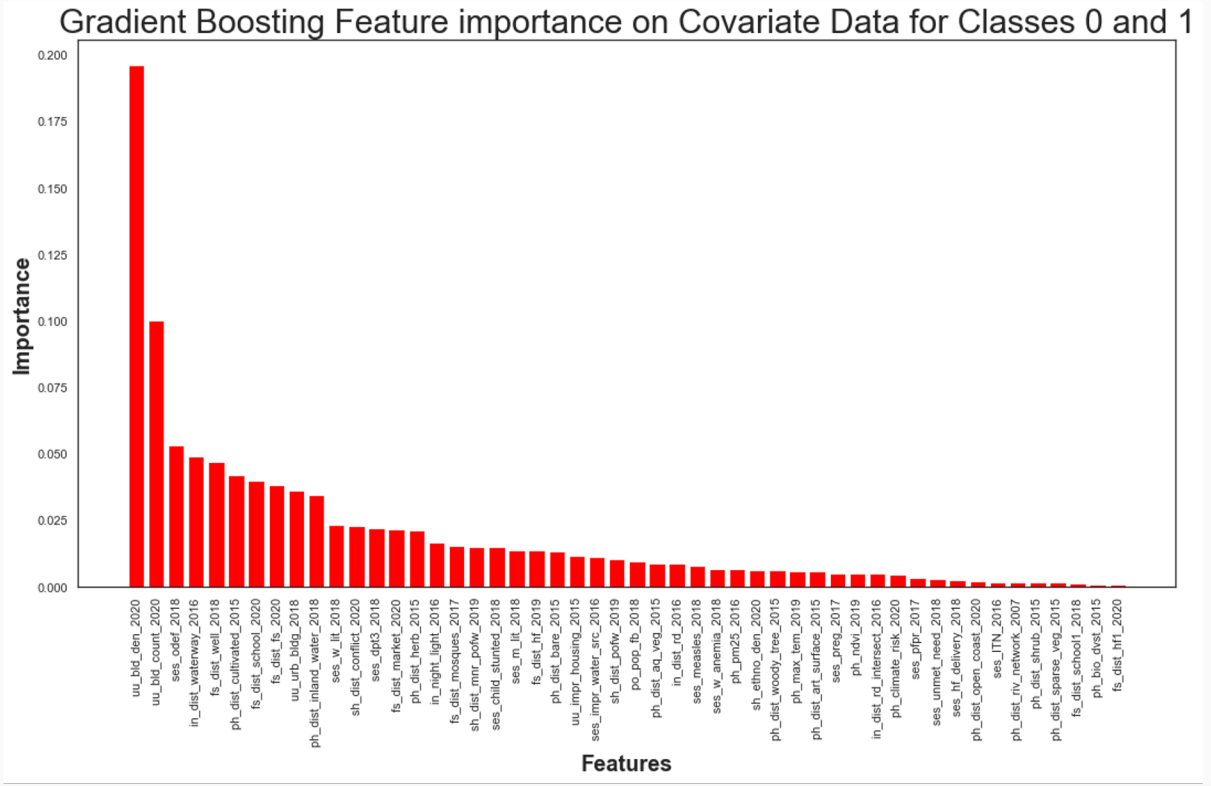
Feature Ranking: Logistic Regression

- Feature importance conducted on Validation set
- Most significant feature:
  - ‘ses_odef_2018”



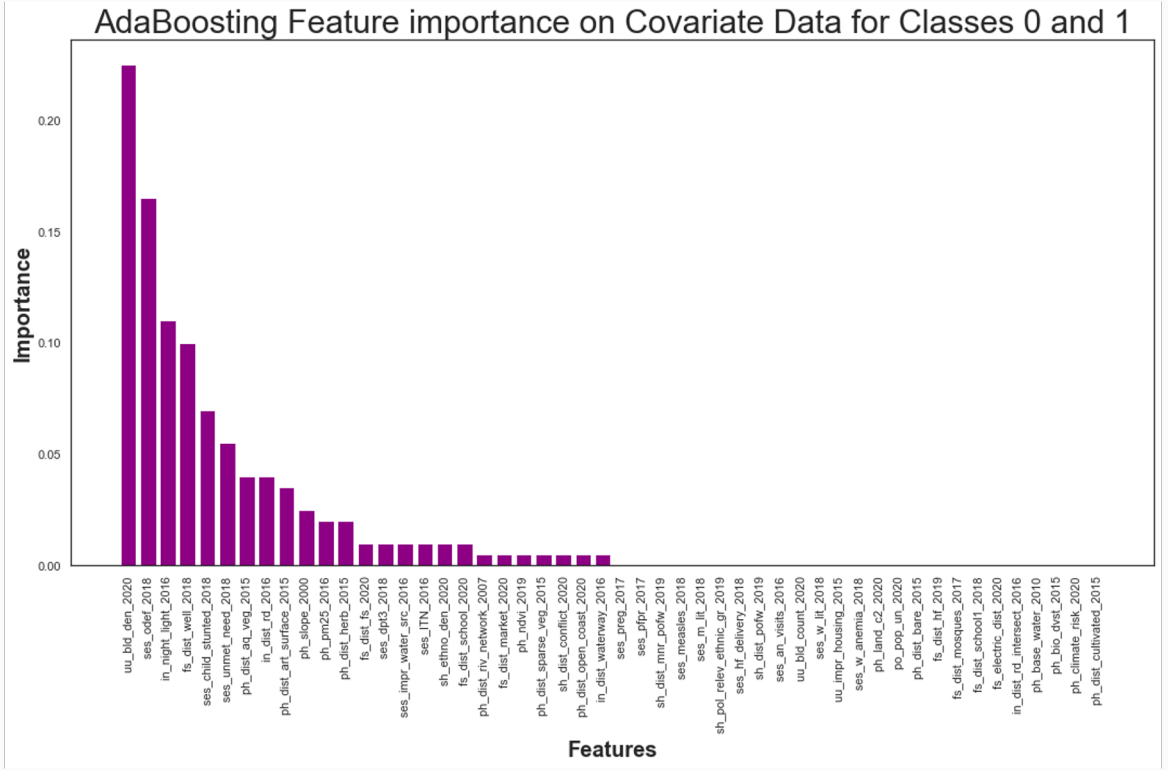
Feature Ranking: Gradient Boosting

- Feature importance conducted on Validation set
- Most significant feature:
  - ‘uu_bld_den_2020’



Feature Ranking: Adaptive Boosting

- Feature importance conducted on Validation set
- If feature was not significant, it was given value of zero
  - 25 features were important
- Most significant feature:
  - ‘uu_bld_den_2020’



Feature Ranking: Results

- **No clear pattern was detected amongst covariate feature methods**
  - **Feature with rank 1: ‘uu_bld_den_2020’**

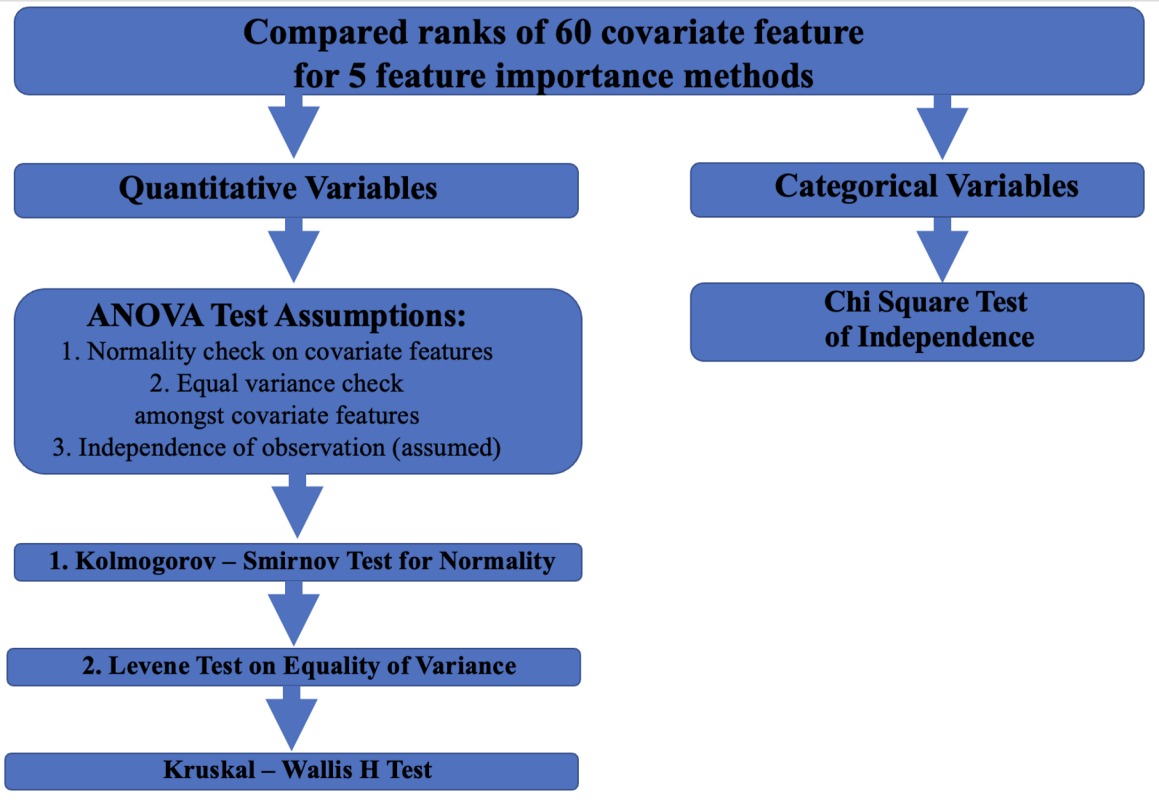
Model	Validation F1 for ‘Deprived’
Random Forest	0.91
Logistic Regression	0.87
Gradient Boosting	0.81
Adaboost	0.79

Feature Ranking: Comparing all Ranks

- Covariate features would have feature importance values for some models but have low values for others
- Top performing feature: ‘uu_bld_den_2020’

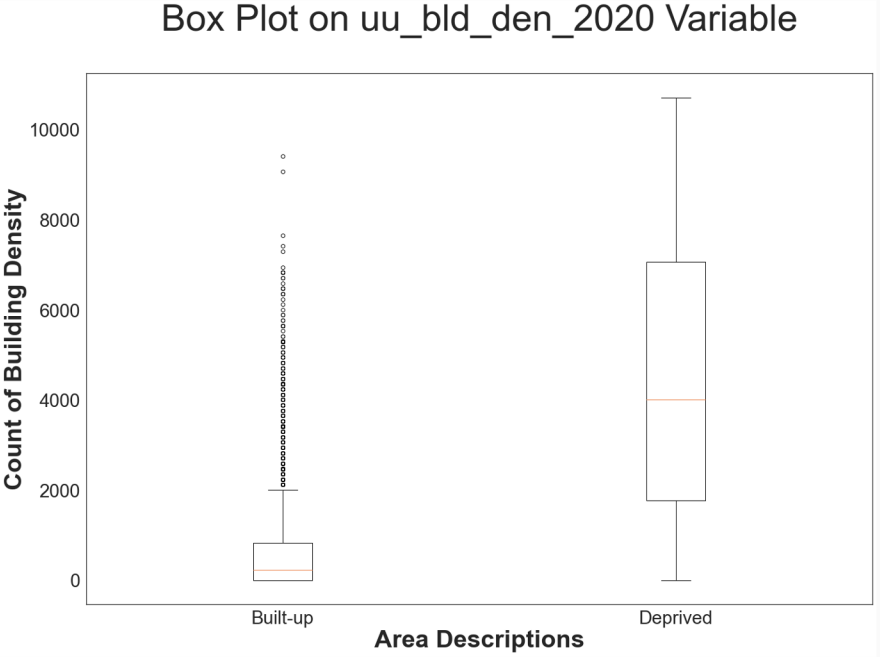
Covariate_ features	top_ logistic_0_1	top_ Random_ Forest_0_1	top_ Gradient_ Boosting_0_1	top_ Ada_ Boosting_0_1	minfo_ 0_1	rank
uu_bld_den_2020	24	1	1	1	4	1
ses_odef_2018	1	13	3	2	14	2
uu_bld_count_2020	25	2	2	35	7	3
ses_impr_water_src_2016	2	9	24	15	21	4
ses_dpt3_2018	18	17	13	14	17	5
ph_dist_ aq_veg_2015	13	5	27	7	35	6

Flow of Covariate Feature Statistical Analysis



Quantitative Statistical Analysis

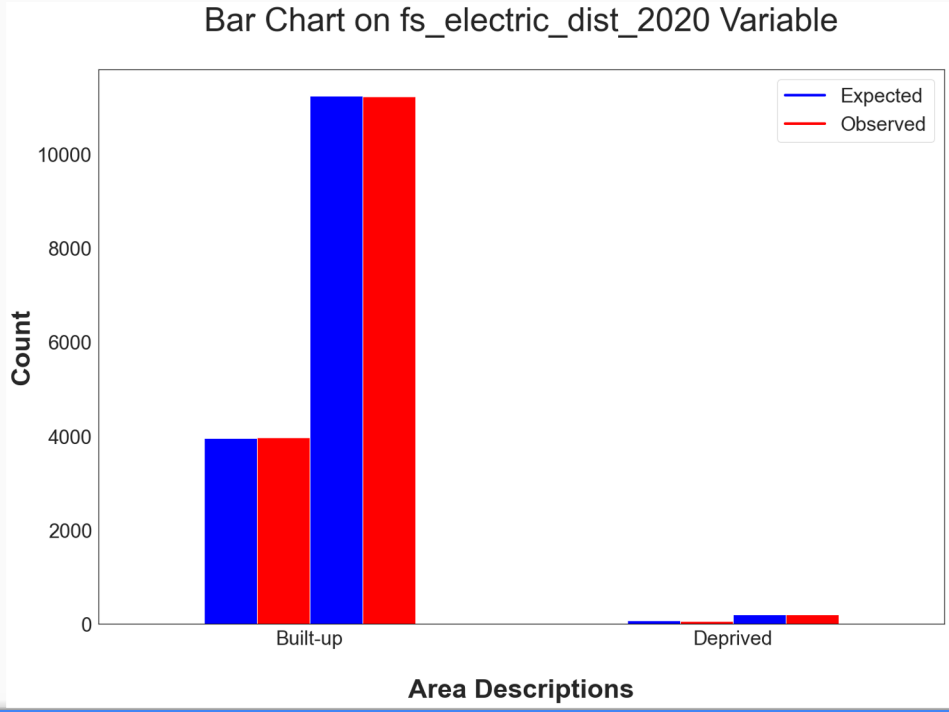
- **Results: Determined that top 21 covariate features were statistically significant in determining ‘Deprived’ and ‘Built-up’ areas**
- Hypothesis testing:
  - **Kolmogorov-Smirnov** test on Normality (ANOVA assumption 1)
  - **Levene** test for equality of variances (ANOVA assumption 2)
  - **Kruskal-Wallis H test**
    - Mean rank utilized for testing (link)
    - Non-parametric rank based testing
- Box plot:
  - Quantiles for ‘Deprived’ and ‘Built-up’ do not overlap (visual indication of difference between groups)



Categorical Statistical Analysis

- **Determined that 1 out of 5 variables was statistically independent from being able to classify 'Deprived' and 'Built-up'**
  - 'fs_electric_dist_2020'
- Hypothesis testing:
  - Chi Square test of Independence

Chi-Square Test Results		
Categorical Covariate Features	**P-value**	**Number of categories**
fs_electric_dist_2020	0.7381	2
ph_hzd_index_2011	0.00	6
ph_land_c1_2019	0.00	11
ph_land_c2_2020	0.00	8
sh_pol_relev_ethnic_gr_2019	0.00	2
uu_urb_bldg_2018	0.00	3



Model Testing: Covariate - classes 0 & 1

- 4 model types tested:
  - Multilayer Perceptron
  - Random Forest
  - LogisticRegression
  - Gradient Boosting
- train/test/val split: 60/20/20
- GridsearchCV used for hyperparameter tuning
- Top 50 features in each feature reduction set tested
- Best performing model:
  - MLP
  - Using all features available

Dataset	Model	Feature Set	Feature Count	Classes	F1 - Class 1 (Deprived)	F1 - Macro
covariate	MLP	All_Features	60	classes_0&1	0.96	0.98
covariate	MLP	Gradient_Boosting_Features	50	classes_0&1	0.96	0.98
covariate	MLP	Logistic_Features	50	classes_0&1	0.94	0.97
covariate	Random_Forest	ADA_Features	50	classes_0&1	0.94	0.97
covariate	MLP	ADA_Features	50	classes_0&1	0.94	0.97
covariate	MLP	Minfo_Features	50	classes_0&1	0.93	0.96
covariate	Logistic_Regression	ADA_Features	50	classes_0&1	0.92	0.96
covariate	Random_Forest	Minfo_Features	50	classes_0&1	0.92	0.96
covariate	Random_Forest	All_Features	60	classes_0&1	0.92	0.96
covariate	Random_Forest	Gradient_Boosting_Features	50	classes_0&1	0.92	0.96
covariate	Random_Forest	Logistic_Features	50	classes_0&1	0.91	0.95
covariate	Logistic_Regression	Gradient_Boosting_Features	50	classes_0&1	0.90	0.95
covariate	Logistic_Regression	Minfo_Features	50	classes_0&1	0.90	0.95
covariate	Logistic_Regression	All_Features	60	classes_0&1	0.89	0.95
covariate	Gradient_Boosting	ADA_Features	50	classes_0&1	0.89	0.94
covariate	Gradient_Boosting	All_Features	60	classes_0&1	0.89	0.94
covariate	Gradient_Boosting	Gradient_Boosting_Features	50	classes_0&1	0.88	0.94
covariate	Gradient_Boosting	Minfo_Features	50	classes_0&1	0.86	0.93
covariate	Gradient_Boosting	Logistic_Features	50	classes_0&1	0.85	0.92
covariate	Logistic_Regression	Logistic_Features	50	classes_0&1	0.82	0.91

Best Model: Covariate - classes 0 & 1

Best Performing Covariate Model on Classes 0 & 1

Dataset	*Model*	*Feature Set*	*Feature Count*	*Classes*	*F1-Deprived*	*F1-Macro*
covariate	*MLP*	*All*	*144*	*classes 0&1*	*0.96*	*0.98*

Covariate Confusion Matrix - Classes 0 & 1

True	***Built-up***	*3040*	*3*
	Deprived	*1*	*50*
		Built-up	***Deprived***
	Predicted		

(Confusion Matrix for the best performing model trained on the covariate features dataset containing only instances of the built-up (0) and deprived (1) classes.)

Model Parameters

- Hidden layer sizes: (100, 100, 100)
  - Number of neurons in hidden layer
- Activation: tanh
  - the hyperbolic tan function, returns f(x) = tanh(x)
- Solver: adam
  - refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba
- Alpha: 0.0001
  - L2 penalty (regularization term) parameter
- Learning Rate: invscaling
  - gradually decreases the learning rate at each time step ‘t’ using an inverse scaling exponent of ‘power_t’. effective_learning_rate = learning_rate_init / pow(t, power_t)

Model Testing: Covariate - all classes

- 4 model types tested:
  - Multilayer Perceptron
  - Random Forest
  - LogisticRegression
  - Gradient Boosting
- train/test/val split: 60/20/20
- GridsearchCV used for hyperparameter tuning
- Top 50 features in each feature reduction set tested
- Best performing model:
  - Random Forest
  - Top 50 Mutual information features

Dataset	Model	Feature Set	Feature Count	Classes	F1 - Class 1 (Deprived)	F1 - Macro
covariate	Random_Forest	Minfo_Features	50	all_classes	0.94	0.98
covariate	Random_Forest	All_Features	60	all_classes	0.91	0.97
covariate	Random_Forest	Gradient_Boosting_Features	50	all_classes	0.91	0.97
covariate	Random_Forest	Logistic_Features	50	all_classes	0.91	0.96
covariate	Random_Forest	ADA_Features	50	all_classes	0.90	0.96
covariate	MLP	Gradient_Boosting_Features	50	all_classes	0.90	0.96
covariate	MLP	ADA_Features	50	all_classes	0.89	0.96
covariate	MLP	Logistic_Features	50	all_classes	0.89	0.96
covariate	MLP	All_Features	60	all_classes	0.88	0.96
covariate	MLP	Minfo_Features	50	all_classes	0.88	0.96
covariate	Logistic_Regression	Logistic_Features	50	all_classes	0.80	0.92
covariate	Logistic_Regression	All_Features	60	all_classes	0.78	0.91
covariate	Logistic_Regression	Gradient_Boosting_Features	50	all_classes	0.78	0.91
covariate	Gradient_Boosting	ADA_Features	50	all_classes	0.78	0.92
covariate	Gradient_Boosting	Logistic_Features	50	all_classes	0.78	0.92
covariate	Logistic_Regression	ADA_Features	50	all_classes	0.78	0.91
covariate	Logistic_Regression	Minfo_Features	50	all_classes	0.78	0.90
covariate	Gradient_Boosting	Gradient_Boosting_Features	50	all_classes	0.74	0.91
covariate	Gradient_Boosting	All_Features	60	all_classes	0.73	0.90
covariate	Gradient_Boosting	Minfo_Features	50	all_classes	0.71	0.90

Best Model: Covariate - all classes

Best Performing Covariate Model on All Classes

Dataset	Model	Feature Set	Feature Count	Classes	F1-Deprived	F1-Macro
covariate	RF	Minfo	144	all classes	0.94	0.98

Covariate Confusion Matrix - All Classes

True	Built-up	3059	0	11
	Deprived	2	46	4
	Non-built-up	18	0	5975
		Built-up	Deprived	Non-built-up
	Predicted			

(Confusion Matrix for the best performing model trained on the full covariate features dataset containing all instances of the Built-up (0), Deprived (1), and Non-built-up (2) classes)

Conclusion

Table of Contents

Original Raw Image of Lagos, Nigeria

Conclusion: Problem Statement

There were two questions the researchers hoped to address in this study:

1. **Can feature importance methods derive any contextual or covariate features that are useful in identifying 'Deprived' and 'Built-up' areas?**
   a. Feature importance methods based on Mutual information, Random Forest, Gradient Boosting, Logistic regression, and Adaptive boosting were able to identify the ranks of Contextual and Covariate features
      i. Top Covariate feature: 'uu_bld_den_2020'
   b. Proved statistically that top 21 Covariate features were significant
2. **Through the use of classical machine learning models and statistical analysis, are contextual and/or covariate features useful in identifying 'Deprived' and 'Built-up' areas?**
   a. Covariate features were useful
   b. Contextual features were not useful

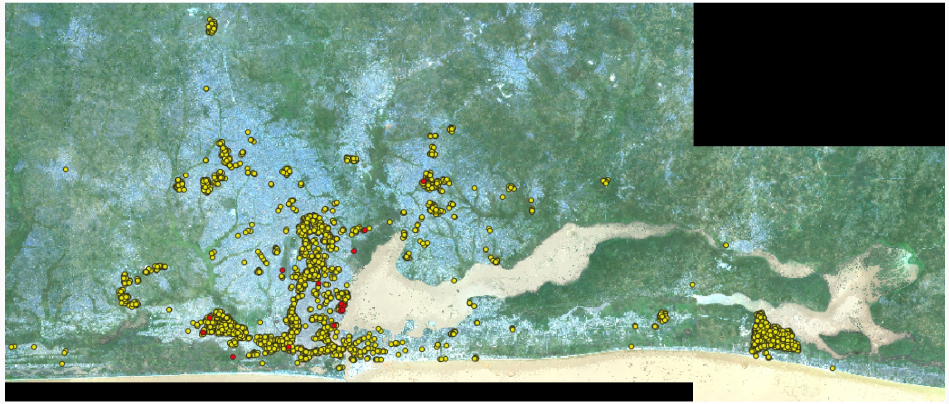
BEST PERFORMING MODELS

Dataset	*Model*	*Feature Set*	*Feature Count*	*Classes*	*F1-Deprived*	*F1-Macro*
covariate	*MLP*	*All*	*60*	*classes 0&1*	*0.96*	*0.98*
covariate	*RF*	*Minfo*	*50*	*all classes*	*0.94*	*0.98*
contextual	*GB*	*All*	*144*	*classes 0&1*	*0.35*	*0.67*
contextual	*MLP*	*All*	*144*	*all classes*	*0.25*	*0.70*

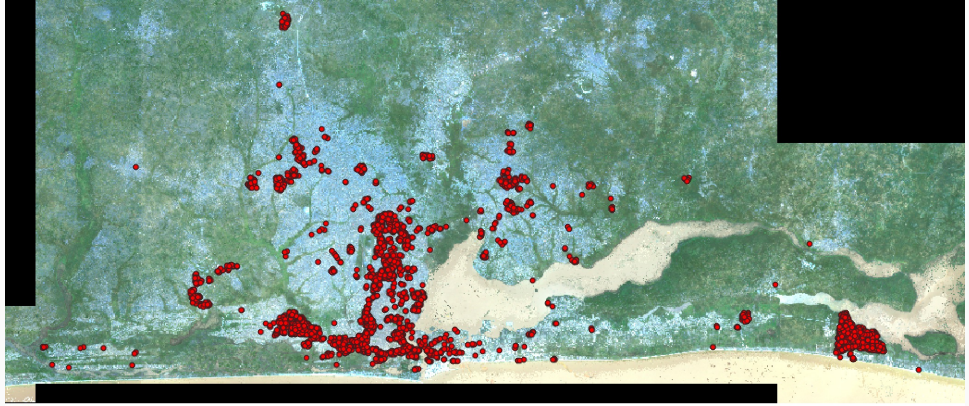
A. MLP: Multilayer Perceptron
B. RF: Random Forest
C. GB: Gradient Boosting

Predicted Deprived -

Predicted Built-up -



*https://unstats.un.org/sdgs/report/2019/goal-11/



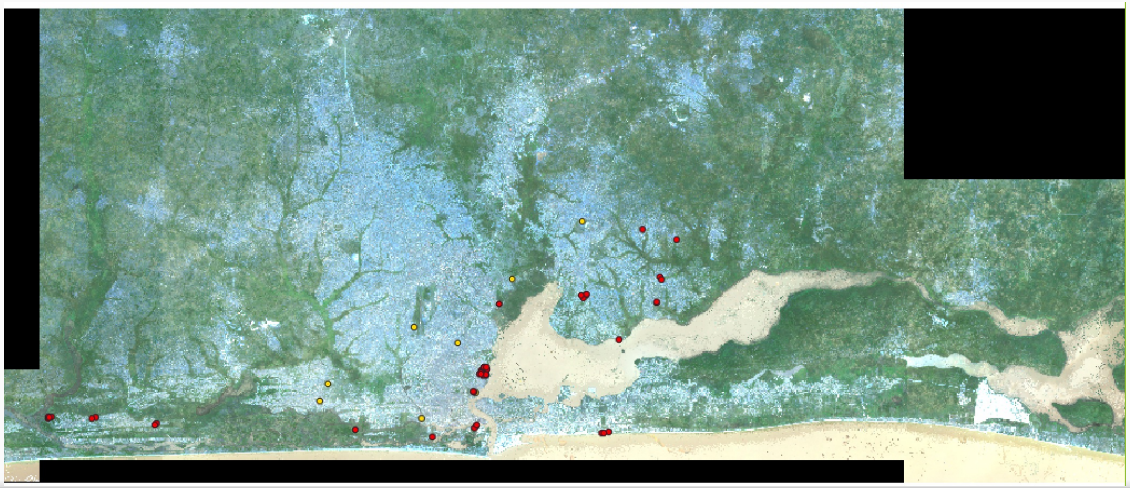
Model mainly predicted all labels as Built-up… that is why no yellow values are visible

Actual Built- up -

Predicted Built-up -

Actual Deprived -

Predicted Deprived -

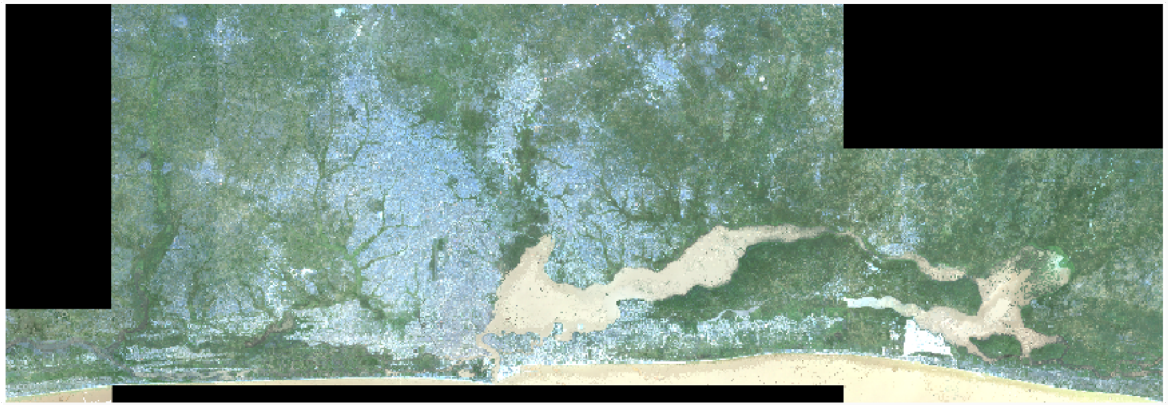


*https://unstats.un.org/sdgs/report/2019/goal-11/

Discussion

Table of Contents

Original Raw Image of Lagos, Nigeria



Discussion

- **Add more ‘Deprived’ areas to analysis to address class imbalance**
- **Attempt different processing steps for extracting the contextual features**
  - **Taking max, min, or median values**
- **Investigate outliers/ apply transformations within ‘Deprived’ and ‘Built-up’ covariate features to further confirm statistical analysis**

*https://unstats.un.org/sdgs/report/2019/goal-11/

Questions?