



Will it rain tomorrow?

Machine Learning Project
Kelvin, Lois, Michael



01

Project Aims

Purpose and business value of
this project

02

Data Preprocessing

EDA, imputation, encoding,
feature scaling

03

Modelling

Logistic regression, random
forest, decision tree, SVM

04

Conclusion

Model evaluation and future
plans

01

Project Aims



Purpose

- Build a binary classification model capable of accurately predicting rainfall
- Determine the best algorithm by evaluating accuracy and variance of different models



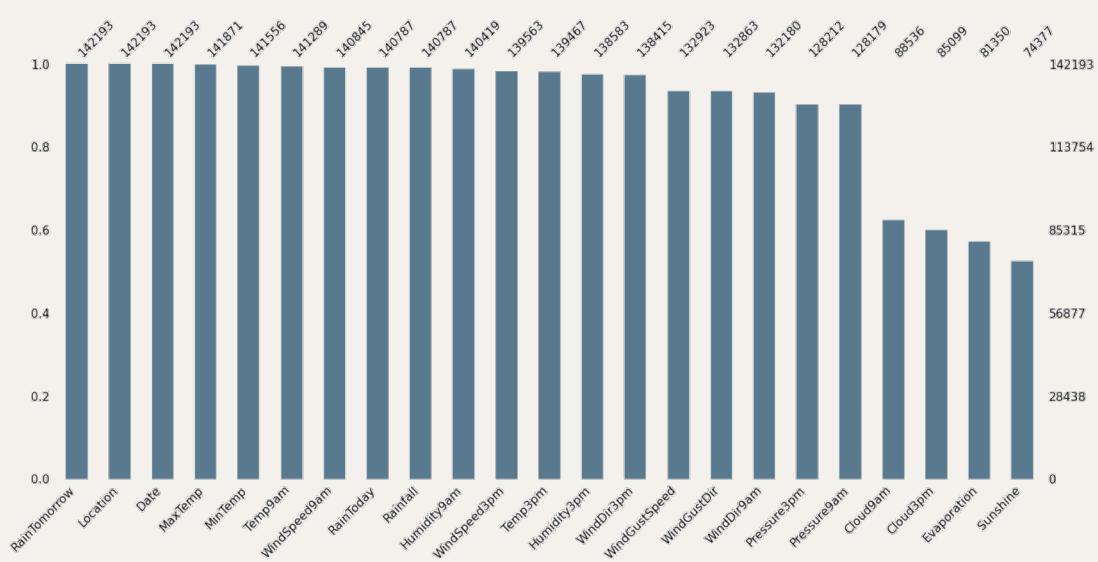
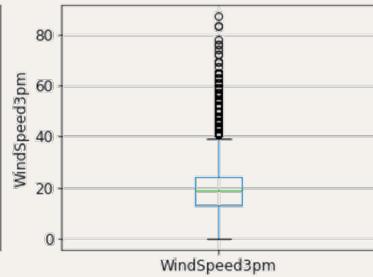
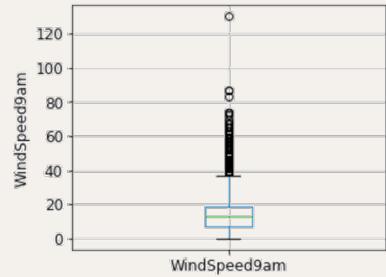
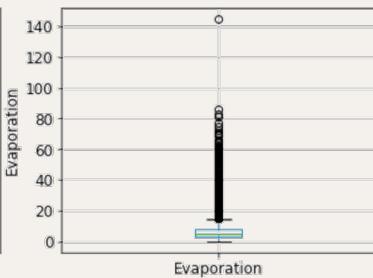
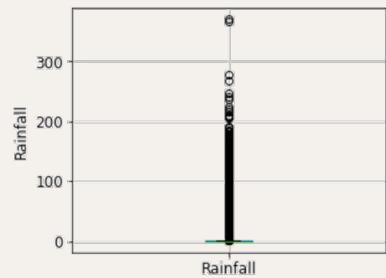
Business Value



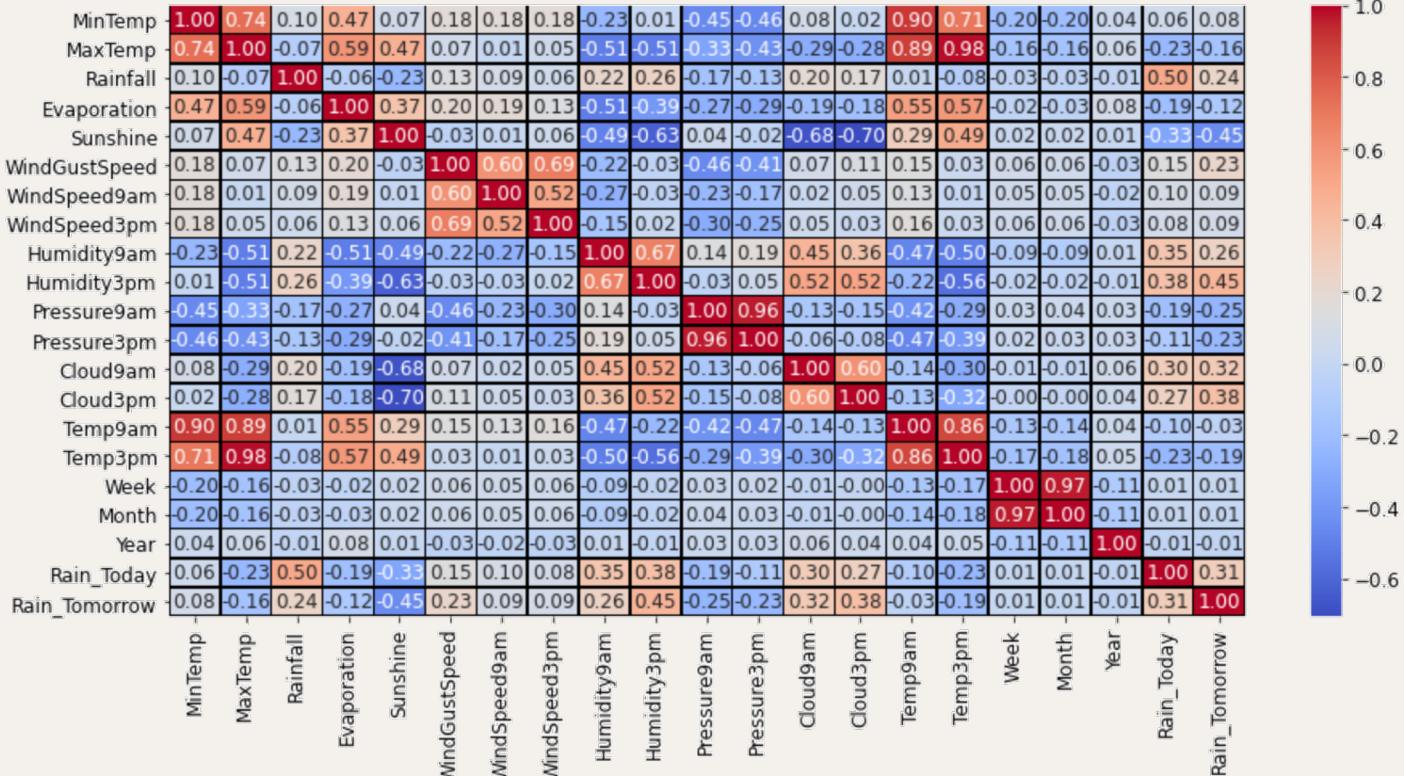
- Aviation, tourism and agriculture industries are especially sensitive to the weather
- Accurate weather forecasting is essential to these businesses as it allows for better preparation

02 Data Preprocessing

Exploratory Data Analysis



Exploratory Data Analysis



Feature Engineering

Data Split

TrainTestSplit before feature engineering to prevent data leakage



Imputation

Missing values imputed with median for numerical data; mode for categorical data

Encoding

Categorical data converted to numerical data using One-hot encoding

Scaling

Standardisation of features using StandardScaler



Dataset

	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm
0	0.458142	0.095453	-0.571161	2.064725	0.569429	0.387781	-1.627919	0.804606	-2.332677	-1.922876
1	-0.387724	-1.268856	2.128528	0.769746	0.115690	1.078135	1.169788	0.220135	-1.399388	-1.129926
2	0.160523	0.306428	-0.571161	-0.753759	-0.905225	-0.187514	0.536345	0.561076	1.119013	0.934739
3	-0.716672	-1.184466	-0.571161	4.121456	1.817213	1.883547	-0.888902	-0.315630	-1.932696	-2.072490
4	0.113530	-0.143653	-0.571161	-0.220532	0.115690	0.157663	0.008476	0.025312	0.793102	0.889855
...
113749	-0.497373	-0.607799	2.128528	0.236519	0.682864	-0.417632	0.061263	0.025312	-0.021674	0.366208
113750	0.144859	-0.467149	1.959797	-0.068182	-1.358964	-0.647749	1.117001	0.171429	0.067210	0.441015
113751	-0.215418	0.320493	-0.571161	0.312695	1.363473	-0.417632	-1.469558	-1.484571	-0.288329	-0.396820
113752	-0.137097	-0.256173	-0.571161	-0.068182	0.569429	0.387781	0.325197	1.340370	-0.006860	-0.007826
113753	-0.293739	0.686184	-0.571161	0.465045	0.115690	-0.647749	-0.941689	-1.143630	-0.006860	-0.007826

03 Modelling

Logistic Regression

LOGISTIC REGRESSION

Model accuracy score: 0.8411

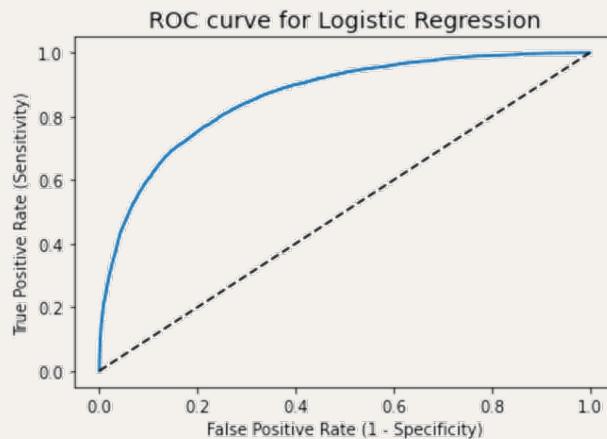
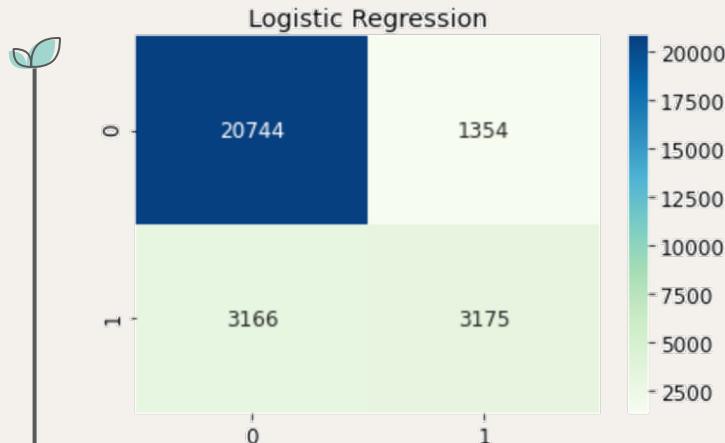
Training set score: 0.8445

Test set score: 0.8411

Time Taken: 1.3230

ROC AUC : 0.8610

	precision	recall	f1-score	support
0	0.87	0.94	0.90	22098
1	0.70	0.50	0.58	6341
accuracy			0.84	28439
macro avg	0.78	0.72	0.74	28439
weighted avg	0.83	0.84	0.83	28439



Random Forest

RANDOM FOREST

Model accuracy score: 0.8237

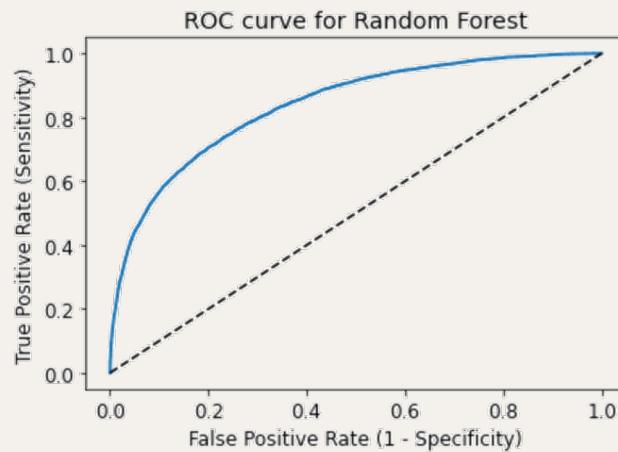
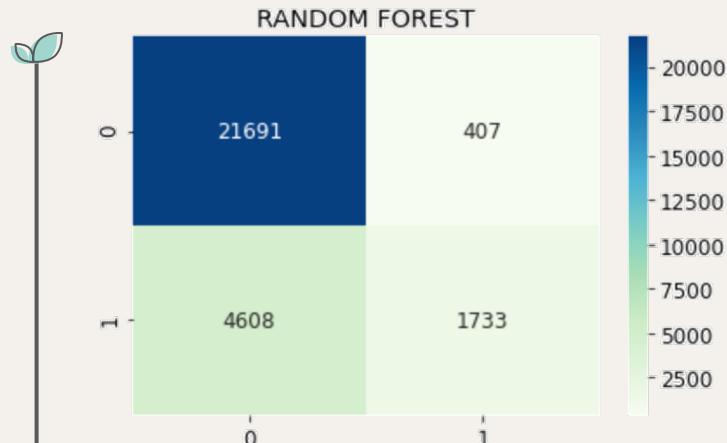
Training set score: 0.8263

Test set score: 0.8237

Time Taken: 36.9789

ROC AUC : 0.8377

	precision	recall	f1-score	support
0	0.82	0.98	0.90	22098
1	0.81	0.27	0.41	6341
accuracy			0.82	28439
macro avg	0.82	0.63	0.65	28439
weighted avg	0.82	0.82	0.79	28439



Support Vector Machines (SVM)

SUPPORT VECTOR MACHINES

Model accuracy score: 0.8407

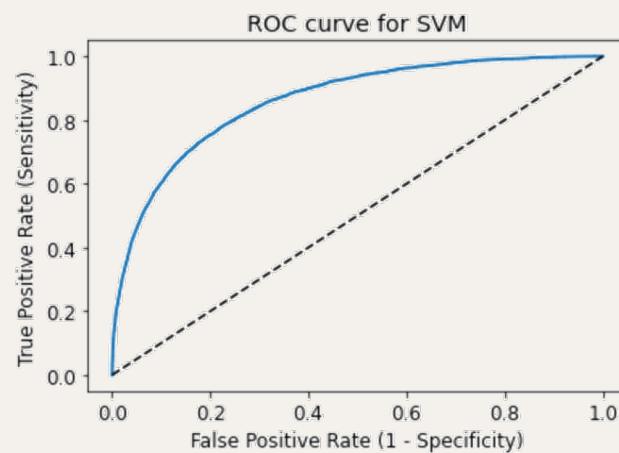
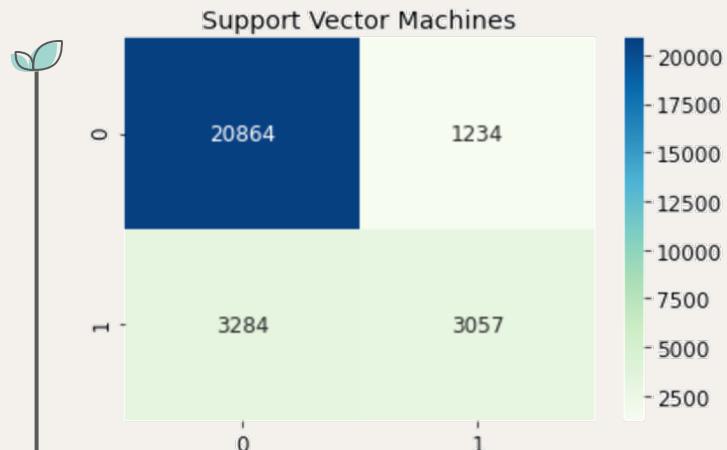
Training set score: 0.8441

Test set score: 0.8407

Time Taken: 71.7474

ROC AUC : 0.7127

	precision	recall	f1-score	support
0	0.86	0.94	0.90	22098
1	0.71	0.48	0.57	6341
accuracy			0.84	28439
macro avg	0.79	0.71	0.74	28439
weighted avg	0.83	0.84	0.83	28439



Decision Tree

DECISION TREE

Model accuracy score: 0.7860

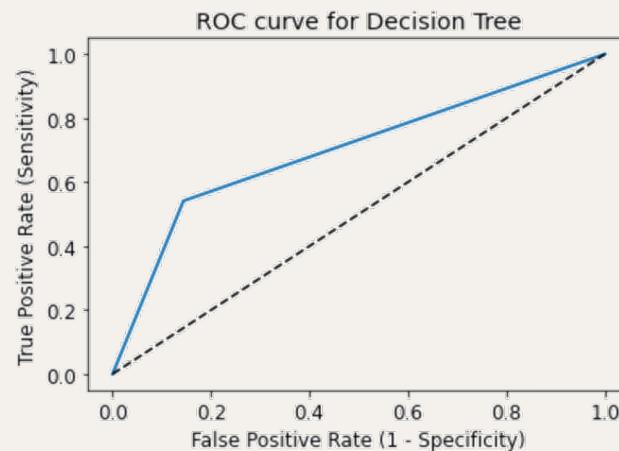
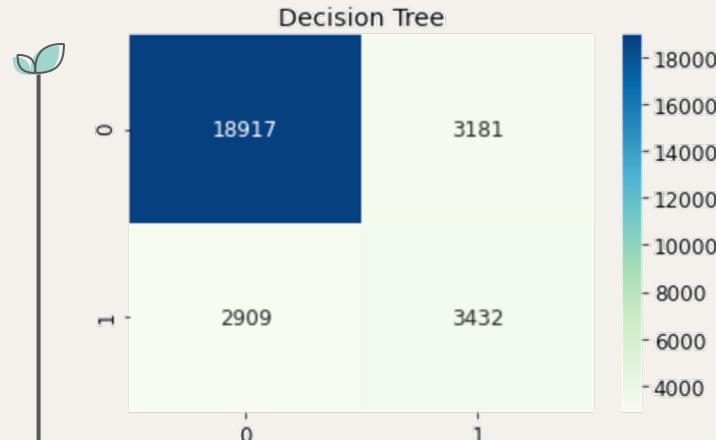
Training set score: 1.0000

Test set score: 0.7860

Time Taken: 6.7330

ROC AUC : 0.6986

	precision	recall	f1-score	support
0	0.87	0.86	0.86	22098
1	0.52	0.54	0.53	6341
accuracy			0.79	28439
macro avg	0.69	0.70	0.70	28439
weighted avg	0.79	0.79	0.79	28439



XGBoost

XGBOOST

Model accuracy score: 0.8467

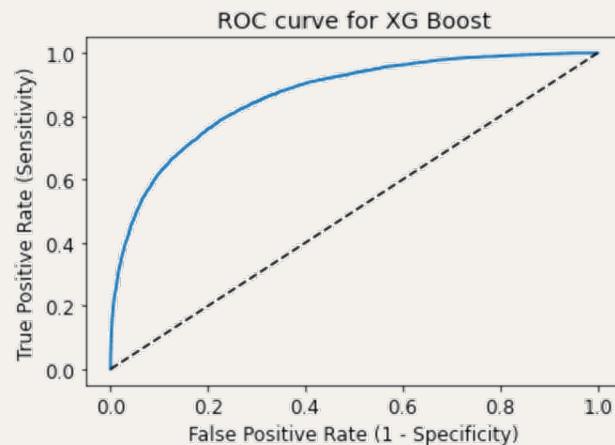
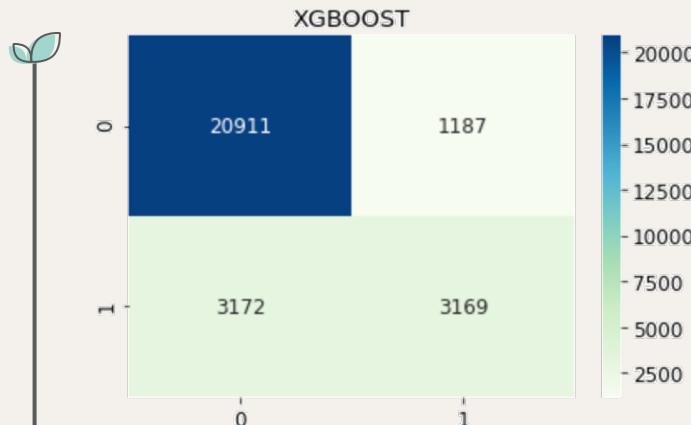
Training set score: 0.8507

Test set score: 0.8467

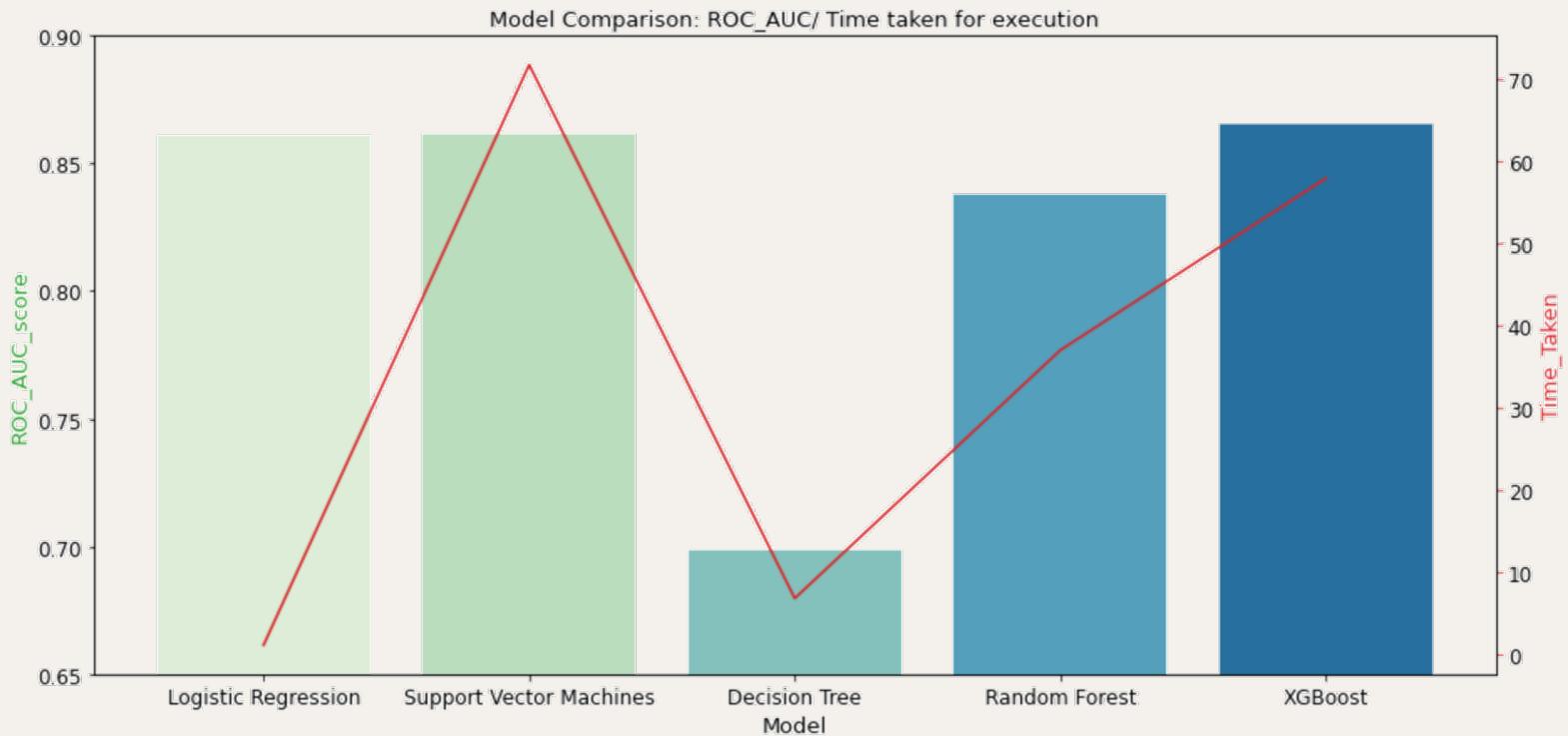
Time Taken: 57.9336

ROC AUC : 0.8653

	precision	recall	f1-score	support
0	0.87	0.95	0.91	22098
1	0.73	0.50	0.59	6341
accuracy			0.85	28439
macro avg	0.80	0.72	0.75	28439
weighted avg	0.84	0.85	0.84	28439



Comparison

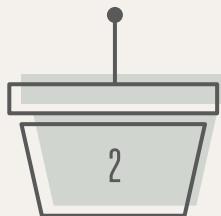


04 Conclusion

Model Evaluation

Logistic Regression

- Accuracy: 0.8411
- ROC-AUC: 0.8610
- Train vs. test score delta: 0.0034
- Time taken: 1.3s



XGBoost

- Accuracy: 0.8467
- ROC-AUC: 0.8653
- Train vs. test score delta: 0.0040
- Time taken: 57.9s

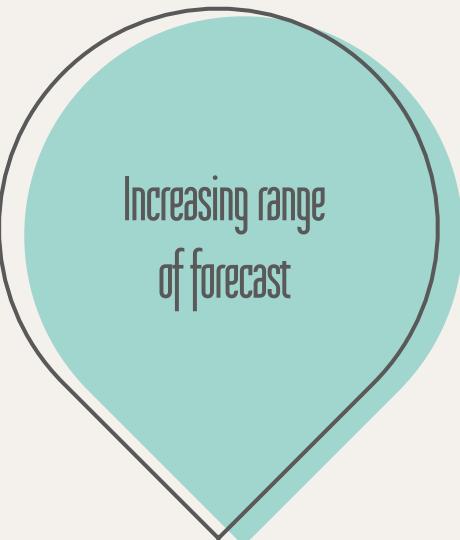


Random Forest

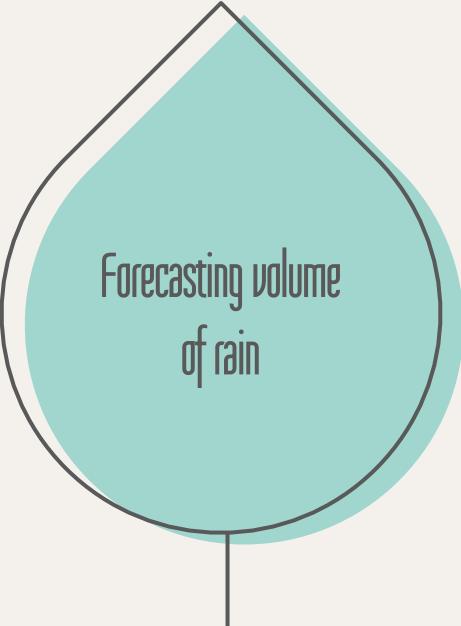
- Accuracy: 0.8237
- ROC-AUC: 0.8377
- Train vs. test score delta: 0.0026
- Time taken: 37.0s



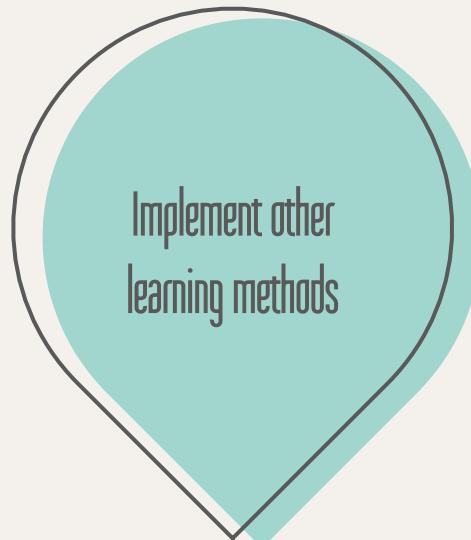
Future Plans



Increasing range
of forecast



Forecasting volume
of rain



Implement other
learning methods

