Dissertation Submitted for the partial fulfilment of the **B.Sc. as a part of M.Sc. (Integrated) Five Years Program AIML/Data Science** degree to the Department of AIML & Data Science.

Project Dissertation

VISIBILITY PREDICTION

Submitted to

By

Yatrik Shah (Roll no. - 40)

Semester-VI

M.Sc. (Integrated) Five Years Program AIML

Department of AIML & Data Science.
School of Emerging Science and Technology
Gujarat University

June 2022

DECLARATION

This is to certify that the research work reported in this dissertation entitled "**VISIBILITY PREDICTION**" for the partial fulfilment of B.Sc. as a part of M.Sc. (Integrated) in Artificial Intelligence and Machine Learning/Data Science degree is the result of investigation done by myself.

Place: Ahmedabad Yatrik Shah

Date: June 7 2022

ACKNOWLEDGEMENT

On this occasion of submitting my project work, I would like to thank all the people who have made this possible. The role of the Department of AIML, Gujarat University in shaping my professional capability is very important.

I thank Dr. Ravi Gor, coordinator of the Department of AIML, Data Science & Actuarial Science, for all his support.

My guide Rashmi Madam has always guided me on the right path and has been a motivating source for innovative project work. This work would have not been possible without her guidance, support, and encouragement. With the help of her guidance, I magnificently overcame the difficulties during my work and learned at each step.

I am heartily and loving thankful to my parents for their sacrifice, support and endless love which encouraged me to efficiently overcome the difficulties in my pursuit of the project and who always helped me in my work.

Last but not the least, I would not be able to complete the work without the blessings of almighty and would like to thank God for giving me strength and patience to carry out my work with full dedication.

-Yatrik Shah

Contents

1. ABSTRACT

In today’s world, the problem of human safety is increasing day by day due to absorption of light to reduce the clarity and colours of what you see. This affects the visibility in the environment, and it goes down. Many factors can cause visibility in the environment like haze, air pollution, climate change etc. Considering this, I have decided to make the machine to predict the level of visibility in the environment taking various features like temperature, wind speed, humidity etc. It becomes very important in cities like Delhi, where the air pollution remains at peak level, or the cities where there is high amount of haze which can cause the air traffic problem. My model will predict the level of visibility level in the environment so we can avoid these problems before it arises.

2. INTRODUCTION

Visibility is a measure of the distance at which an object or light can be clearly perceived. We predict the visibility of an object from a particular distance in kilometers based on: Dry Bulb Temp, Wet Bulb Temp, Relative humidity and many more. Visibility is affected by certain environmental factors and may vary according to the direction and angle of view and the height of the observer. In some certain situations Visibility as a limitation factor affects a lot. Football match, f1 racing, takeoff and landing in Air Traffic Control are some of the examples of these situations where the factor visibility affects a lot. There visibility plays a key role in the public safety factor.

3. REAL LIFE APPLICATIONS

- Key value in deciding if the flight would take off or land or not.
- Flight's Visibility is an important factor in all phases of flight, but especially when an aircraft is maneuvering on or close to ground, at that time weather must be clear in order to avoid accidents.
- In addition, it has very important value in knowing distance from the land while docking a ship at a dockyard.
- Car Racing is not allowed if there isn't proper visibility.
- Useful in day-to-day life as well while driving vehicles.

4. Data description

1. VISIBILITY - Distance from which an object can be seen.
2. DRYBULBTEMPF-Dry bulb temperature (degrees Fahrenheit). Most reported standard temperature.
3. WETBULBTEMPF-Wet bulb temperature (degrees Fahrenheit).
4. DewPointTempF-Dew point temperature (degrees Fahrenheit).
5. Relative Humidity-Relative humidity (percent).
6. Wind-Speed-Wind speed (miles per hour).

7. WindDirection-Wind direction from true north using compass directions.
8. StationPressure-Atmospheric pressure (inches of Mercury; or ‘in Hg’).
9. SeaLevelPressure- Sea level pressure (in Hg).
10. PreciP Total-precipitation in the past hour (in inches).

5. WORKFLOW STEPS

- Raw Data Collection
- Data Pre-processing
- Eda and Data Visualization
- Feature Engineering
- Logging
- Maintain Training Data in Database
- Cross Validation and Hyper Parameter Tuning
- Model Selection
  Deployment
- Maintaining Prediction Database
- Creating API
- Integrating with WEB FRAMEWORK - flask
- Model Deployment

6. Important Tools used in the Project

- Python – 3.9
- Numpy

- Pandas
- Matplotlib
- Sklearn
- Scipy
- Logging
- Flask
- Jinja2
- XGBoost
- Sqlite3
- Joblib
- HTML
- CSS
- Gunicorn
- Heroku

7. METHODOLOGY

7.1 DATA PRE-PROCESSING AND EDA

7.1.1 Observation:

As a first step, the data was observed, and all the features were taken into consideration.

7.1.2 Null Values:

In the dataset, it was checked whether there are null values or not and if it is, how many records it contains.

7.1.3 Trends and Patterns:

Visualized some important graphs and charts to observe the considerable trends and some patterns to know the relationship between the features.

7.2 FEATURE ENGINEERING

7.2.1 Feature Selection:

Basically, Feature selection is considered as the first step in Feature Engineering. Here all the features are selected which are truly helpful in achieving Machine Learning end-goal i.e.(here) Regression.

There are different methods in feature selection to select important features. Some of these are: Chi^2 test, feature importance (by the model), Correlation Matrix Heatmap (using Karl Pearson's correlation and Spearman's Rank correlation), Gini impurity, Entropy and Information Gain, etc.

Here the Correlation Matrix Heatmap is used to select the best features for the machine learning model furthermore, Gini impurity is used in the Random Forest model as attribute selection method.

7.2.2 Scaling:

In Machine Learning, scaling is a very useful step of Feature Engineering. The goal of feature scaling in machine learning is to scale all the important features into same range or scale. Some Machine Learning models expect all the features to be scaled properly [Generally in Standard Normal distribution].

Standardization was used as a method of scaling to bring down all the features to a common scale without distorting the differences in the range of the values.

7.2.3 Multicollinearity:

In Machine Learning Multi-collinearity is the problem where two or
More independent columns or features are correlated internally. It may Cause. In Regression Analysis it is assumed that there is no exact relationship among the exploratory variables and in regression analysis this assumption is violated then the problem of multi-collinearity occurs. Removed multi-collinearity by creating heatmap of correlation metrics.

7.2.4 Imputation [Handling null values]:

In Feature Engineering one of the important steps is considered as Imputation or handling the Null values which means, observing Null values or empty values(thresholds) and replacing them with a best-fit meaningful substitute.

Null values were critically observed, and it was removed from all the features in the dataset.

7.3 LOGGING

Logging is a very fundamental part of every end-to-end project or software. Logging let us track events when the program gets executed, so that when the code crashes, we can check the logs and identify that what caused it. Very robust software and programs can be created using Logging.

Here, logs are maintained and get informed about each information, warnings, errors in any program, which helps us to track the program. Logging is very useful especially in case of a program crash. It is an essential part of troubleshooting application and infrastructure performance.

7.4 MAINTAINED TRAINING AND PREDICTION DATABASE

A database is defined as a structured set of data which is stored in a computer's memory or some cloud platform. In large and advanced projects storing each data is a challenging task. Which can further be used in some other tasks like maintaining a training data database which can further be divided into good data and bad data by data validation.

After data cleaning, exploratory data analysis and feature engineering, dumping the final data for model training in the database.

This data can directly be taken for model retraining purpose.

Here the **prediction database** is the database where the data is stored which were tested on the deployed API after deployment.

In more advanced data science projects Prediction Database plays more important role mostly in retraining approach.

Here SQLite database is maintained.

7.5 CROSS VALIDATION AND HYPER PARAMETER TUNING

- Cross validation and Hyperparameter tuning is one of the most important steps in Machine Learning, since the performance of the machine learning model is affected upon the hyperparameters those are selected for model training.

- Performed Cross validation and Hyperparameter tuning on all the models and selected the best models that performs best on the validation sets according to the given parameter grid.

- Found and selected the best hyperparameters of all the models for model training on Full data.

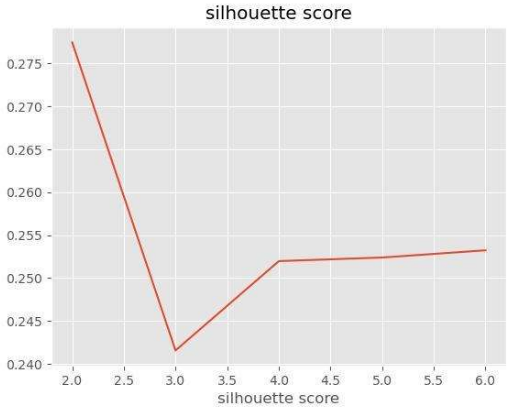
7.5.1 Evaluation Metric

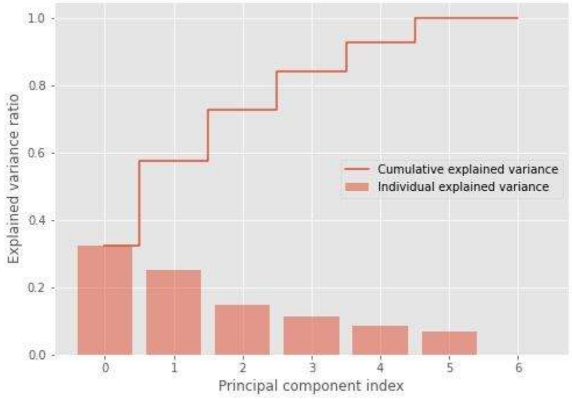
- R Squared or R2 score:
  Here R Squared is used as evaluation metric. R squared is basically a goodness of fit measure for regression models.
- R squared is also called Coefficient of Determination, where R is correlation coefficient. It generally ranges between 0 to 1.
- R squared tends to 1 refers a good model and tends to 0 refers to a bad model.

7.5.2 Clustering Approach

- For better results, the **clustering approach** is used, which means here the best possible clusters are drawn from the data (without considering the target column – based on the input features) and on each cluster, firstly the target is assigned back accordingly. Then by hyper-parameter tuning the best model is selected on those clusters.

- Here K-means clustering is used as the clustering algorithm on the data. Here to cluster the data the necessary steps done are scaling and dimensionality reduction using PCA - Principal component analysis. The optimum value for principal components is achieved by the variance retained method and around 95% variance retained within 4 components. To optimize the value of k in the k-means algorithm, the silhouette score method is used. And the optimized value for the cluster is 2





- And at the time of predictions, it is checked that the input example belongs to which cluster then its prediction is done using that respective model only, which tends to get the best results possible because every cluster has different types of points (which have dis-similarities). So, it may not be possible for a single model to understand all these patterns.
- This is how the clustering approach is used in regression problems.

7.5.3 Used Models:

7.5.3.1 RandomForest :

- RandomForest is an ensemble learning algorithm. It is a bagging or bootstrap model. The base estimator of Randomforest is Decision Tree.
- Basically Bootstrap means, sampling with replacement. Decision trees have high variance, and we can achieve low variance overall by combining multiple decision trees. The concept of RandomForest is that of getting multiple samples from a population data with replacement and training a decision tree model on that data and then combining all those trees and making a single robust model, that is Randomforest.
- In Randomforest regressor the final output is, the average or arithmetic mean of all the outputs of different decision trees, trained on different samples, on a single input. As well as in Randomforesr classifier the mode of the outputs of different decision trees, trained on different samples are used as the final output.

7.5.3.2 XGBoost :

- XGBoost stands for Extreme Gradient Boosting, is also a decision tree-based ensemble machine learning algorithm that fundamentally uses a gradient boosting approach.

- Gradient Boosting is a special case of boosting where gradient descent is used as an optimizer algorithm. The implementation of the Extreme Gradient boosting algorithm is designed for efficiency of computation time and memory resources as it is designed for hardware and software optimization techniques, which gives superior results using less computation time and also fewer computing resources.

XGBoost ensembles tree methods that apply principles of boosting weak learners (Decision Trees) using gradient Descent architecture. The cost function is a combination of losses and regularization terms that prevent overfitting and give the best results.

7.6 Model Selection

- As the cross validation and hyperparameter tuning done we have got the best parameters for the two models Randomforest and XGBoost.

- Results:

  - **Cluster0 XGBoost :**

    XGBoost Best Parameters:

```
{'gamma': 0.4, 'learning_rate': 0.1, 'min_child_weight': 10}
```

XGBoost Test R Squared:

```
0.779091270788543
```

- **Cluster0 Randomforest :**

RandomForest Best Parameters:

```
{'criterion': 'squared_error', 'n_estimators': 400}
```

RandomForest Best Estimators:

```
RandomForestRegressor(n_estimators=400)
```

RandomForest Test R Squared:

```
0.7814797668769428
```

- **Cluster1 XGBoost :**

XGBoost Best Parameters:

```
{'gamma': 0.1, 'learning_rate': 0.17, 'min_child_weight': 5}
```

XGBoost Test R Squared:

```
0.5655753425170551
```

- **Cluster1 Randomforest :**

RandomForest Best Parameters:

{'criterion': 'squared_error', 'n_estimators': 200}

RandomForest Best Estimators:

RandomForestRegressor(n_estimators=200)

RandomForest Test R Squared:

0.5878544038468825

- **FullData XGBoost :**

XGBoost Best Parameters:

{'gamma': 0.4, 'learning_rate': 0.17, 'min_child_weight': 10}

XGBoost Test R Squared:

0.6168362609976628

- **FullData Randomforest :**

RandomForest Best Parameters:

{'criterion': 'squared_error', 'n_estimators': 400}

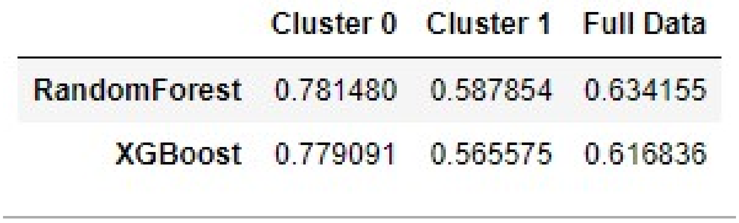
RandomForest Best Estimators:

```
RandomForestRegressor(n_estimators=400)
```

RandomForest Test R Squared:

0.6341548845912297

Summarizing:



	Cluster 0	Cluster 1	Full Data
RandomForest	0.781480	0.587854	0.634155
XGBoost	0.779091	0.565575	0.616836

Here, we can clearly notice that Random Forest model is giving better output on both the respective clusters compared to XGBoost model.

So that the Randomforest model is chosen for both the clusters. In prediction if there is a data point to predict, then first check that data point belongs to which cluster then predict by that model respectively.

8. Deployment

8.1 Creating API

- A real-world Data Science based does not end just after successfully building a machine learning model that performs well on the test dataset.
- In real-world industrial end-to-end Artificial Intelligence or Data Science based projects creating Front-End API is also having equal importance as back-end tasks like- Model Training or Model Evaluation

- The Front-End is created with the help of HTML (Hypertext Markup Language) and CSS (Cascading Style Sheets), which are very powerful tools.
- Flask is used as WEB – API in python that integrates back-end and front-end.
- Flask depends on the Jinja template engine and the Werkzeug WSGI toolkit.

8.2 Integrating with Web-Framework Flask:

Creating API is not enough. It must be connected with some back-end code.

Here Flask framework is used. Python provides a micro web framework named, Flask. Flask is an application framework written in python.

Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine.

Here Flask is used to integrate frontend written in HTML and CSS.

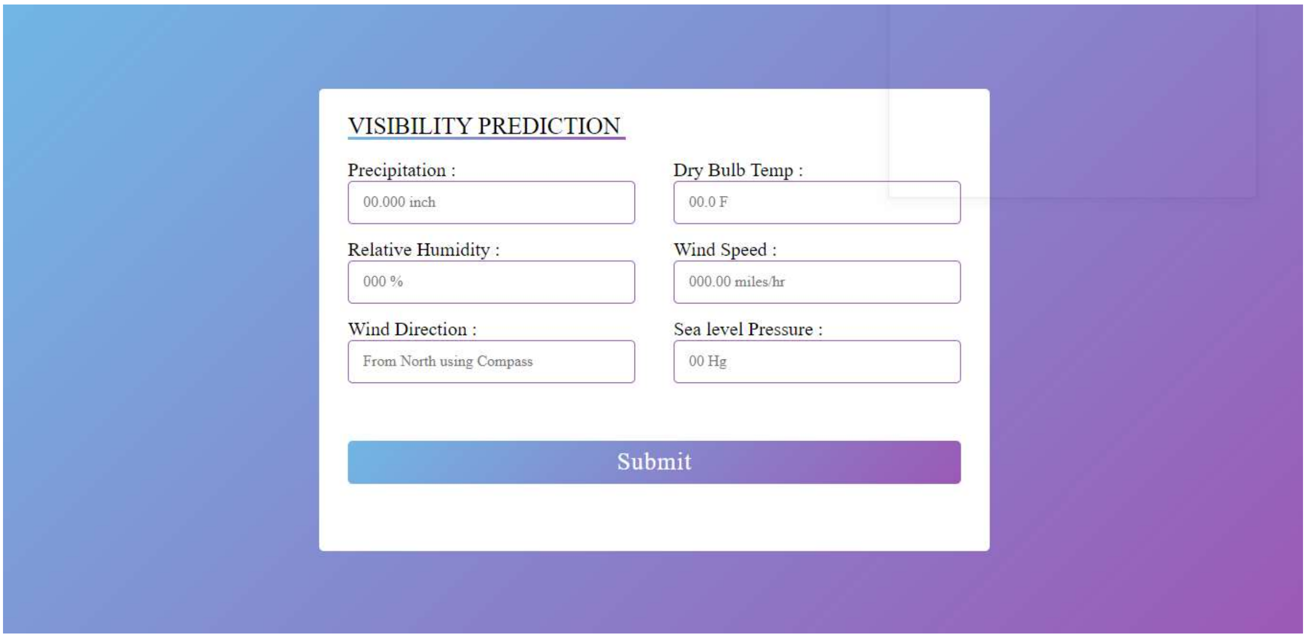
8. 3 MODEL DEPLOYMENT

- Generally Model Deployment is considered as the last step in any Machine Learning project.
- After successfully creating front-end API with HTML and CSS and using Flask as web Framework only Model Deployment is left.
- Here HEROKU platform is chosen as Model Deployment.
- Heroku is a container-based cloud Platform as a Service (PaaS). Developers use Heroku to deploy, manage, and scale modern apps.

Deployment Link:

https://hidden-vision.herokuapp.com/

Glimpse of API



VISIBILITY PREDICTION

Precipitation : 00.000 inch

Dry Bulb Temp : 00.0 F

Relative Humidity : 000 %

Wind Speed : 000.00 miles/hr

Wind Direction : From North using Compass

Sea level Pressure : 00 Hg

9. References

https://en.wikipedia.org/wiki/Visibility

https://www.civilaviation.gov.in/sites/default/files/moca_001421.pdf

https://skybrary.aero/articles/visibility

https://en.wikipedia.org/wiki/Air_traffic_control