

FORECASTING OF INDIAN WEATHER: VISIBILITY PREDICTION USING MACHINE LEARNING TECHNIQUES

Prof. Nikita Ramrakhiani
Assistant Professor- Business Analytics
Vivekanand Business School
Chembur, Mumbai 400-074
nikita.ramrakhiani@ves.ac.in

Mr. Harsh Patel
Student- Business Analytics
Vivekanand Business School
Chembur, Mumbai 400-074
harsh.patel2224p@ves.ac.in

Mr. Amey Rajendra Hulsure
Student- Business Analytics
Vivekanand Business School
Chembur, Mumbai 400-074
amey.hulsure2224p@ves.ac.in

Mr. AbhijeetKumar
Srivastava
Student- Business Analytics
Vivekanand Business School
Chembur, Mumbai 400-074
abhijeetkumar.srivastava@ves.ac.in

Abstract—Weather reports are essential for a range of purposes, including pollution control and addressing weather-related issues. They help identify the dispersion of pollutants, enabling effective pollution management. Weather forecasts also aid in preparing for emergencies, optimizing agriculture practices, ensuring safe transportation, and promoting public health. These reports play a vital role in decision-making across sectors, enhancing safety, resilience, and overall well-being. Using machine learning, particularly the logistic regression algorithm, this study demonstrates its efficacy in predicting binary weather events like rain or sunny based on various weather parameters. Through accurate modeling and evaluation, the research highlights the algorithm's potential for practical weather forecasting applications. This research intends to analyze the significance of weather forecasting for advancing our comprehension of Indian weather dynamics and its role in enhancing predictive capabilities to address weather-related challenges effectively. Through our model and analysis, we attempt to predict the visibility index and identify the factors such as pollution, gas emissions, etc which affect visibility in specific regions which impact the number of accidents and mishaps

Keywords— *Weather Forecasting, Air Pollution, Visibility, Machine Learning, Tableau*

I. INTRODUCTION

Visibility plays a pivotal role in our daily lives, influencing everything from our choice of attire to the planning of outdoor activities and even critical decisions in agriculture, transportation, and disaster management. Visibility, in simple terms, is a measure which is used to see how clear is the atmosphere in a particular region. In a country as geographically diverse as India, with its vast expanse and a

wide range of climates. One crucial aspect of weather prediction, especially in densely populated regions like India, is visibility forecasting. The ability to predict visibility conditions accurately is essential for aviation, road

safety, and various other sectors that rely on clear visibility. Visibility plays an important role in not only disaster management, but also in day to day activities of a person, For eg, If a region with low visibility and the airport is situated at such a place where planes cannot land using autopilot or any modern features, then either the pilot will have to land the plane manually or might have to divert to some other airport to land to avert any disaster from happening due to low visibility. This in turn takes a huge toll on not only the airline, but also on the passengers who were scheduled to land at the region. Delhi, which is a highly polluted city in India, can be considered a great example of how pollutants affect along with various weather conditions affect visibility

Usually visibility forecasting has heavily relied on meteorological models and observational data. However, in recent years, the integration of machine learning techniques has revolutionized the way visibility is predicted in a region. Machine learning algorithms have the potential to process vast amounts of data, recognize complex patterns, and improve the accuracy of visibility predictions.

This paper explores the application of machine learning techniques in forecasting Indian weather, with a specific focus on visibility prediction. By harnessing the power of data-driven models, this research aims to enhance the precision of visibility forecasts, thereby contributing to safer

and more efficient daily life operations, including transportation, agriculture, and disaster preparedness.

In this context, this study delves into the challenges and opportunities in forecasting visibility in the diverse climatic regions of India. It also explores the role of machine learning algorithms in improving the accuracy of visibility predictions, drawing from a rich pool of historical weather data and advanced modeling techniques. Through this research, we also plan to predict how visibility impacts the daily productivity and lives of people living in regions with low visibility. Thus, due to low visibility, we will also be able to predict the impact it has on the rate of accidents in the region where visibility index (GB Defra Index) whose scale ranges from 1- 10 where 1 being the lowest (meaning it has the highest visibility) and 10 being the highest (meaning it has the lowest visibility)

II. RESEARCH METHODOLOGY

Forecasting visibility in India, like in any other region, requires a structured research methodology that incorporates various factors contributing to visibility conditions. Visibility forecasting typically relies on meteorological data, pollution levels, topography, and other relevant parameters. Since we want to find how visibility is impacted by other factors such as pollution level, weather conditions, geographical regions, etc. The primary objective of this research is to predict visibility and examine its fluctuations under different conditions such as pollution levels, weather patterns, and other external factors. The study aims to discern the extent to which these conditions affect visibility within specific regions. Additionally, the research investigates the impact of visibility on the overall living conditions of the residents and evaluates whether it correlates with the rate of accidents. This analysis differentiates regions with low visibility from those with high visibility, providing valuable insights into the relationship between visibility, environmental factors, and societal well-being. Along with that, how visibility impacts the living conditions of the people and does it influence the rate of accidents in regions where visibility is low as compared to regions where visibility is quite high. The dataset used for prediction and evaluation is of India Meteorological Department (IMD). Using tools like Python and Tableau, Exploratory Data Analysis was performed, data cleaning and preprocessing was done according to the EDA and for preparation of the data for model building. Similarly, R too was used to clean and drop certain values which if included, would have interfered with the prediction and analysis. Tableau was used to visualize the data and mapped it on graph using and lastly predicted visibility using machine learning algorithms. Algorithm which was used to predict the visibility was logistic regression.

A. Data cleaning

To get a proper understanding of the data, we studied the variables in the dataset, its data types, its range and what

necessary measures needed to be taken. For this, we used MS Excel to access the dataset. Once we understood the data, we then went ahead and imported the dataset to python for further data cleaning which is a very essential part of Exploratory Data Analysis (EDA). Firstly we have a consolidated look of the columns and its first ten variables and then we studied the variables in it. Since there were columns with missing values and they didn't have any use in our analysis, we decided to drop them entirely. We then studied the data more and found out that there were columns which were not required for analysis. For example, there were columns which had longitude and latitude which had no use in our analysis, so we decided to drop them completely. Secondly, there were many non numerical values which did not influence the analysis if they weren't included in the EDA and predictive model. Hence, we decided to drop such columns as well.

Further, we went on to analyze rows and in condition_text, we found certain variables which were not required in analysis. Hence, we imported the entire dataset into R since manipulating rows of the dataset was quite efficient. Once all the unnecessary variables were removed, we downloaded the dataset and reimported the new dataset to Python.

We converted all the categorical variables into numerical values so that it can be used to perform EDA and build an accurate predictive model.

B. Data Exploration

In this section, we used tableau and python to visualize graphs

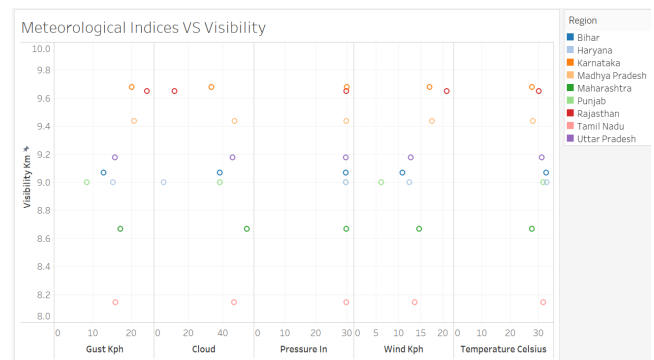


Fig. 1 Visibility with respect to Meteorological Indices

In this graph. We compared meteorological indices like temperature, cloud density, pressure and wind speed with visibility and see whether it influences visibility in any kind or not. Over here, we observe it does not have much influence on visibility.

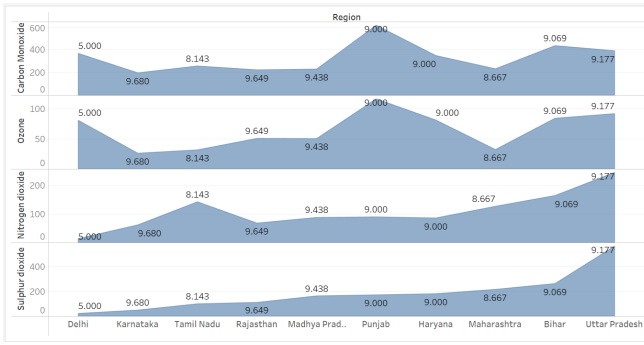


Fig 2. Visibility with respect to Gases

Here, we visualize whether gasses emitted in these regions influence the visibility. The impact gasses do not influence visibility in much way.

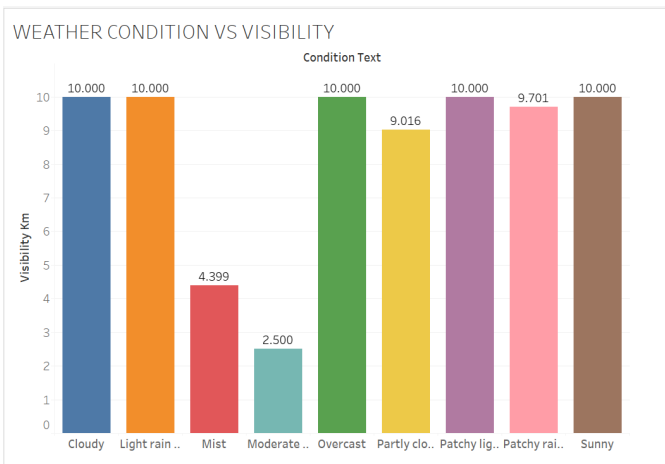


Fig 3. Weather condition VS Visibility

Over here, we compared weather conditions with visibility to see if they influence the visibility in any way. We observe that areas with moderate rainfall have the lowest visibility and regions which have mist also have an average visibility of 4.39 kms whereas cloudy sunny and patchy rainy areas have the highest visibility. We can say that weather conditions greatly influence visibility of a region.

IV. RESULTS

After performing Data transformation and EDA , we started implementing the ML model on model. The Machine Learning Algorithm we used was Logistic Regression.

A. Implementation of the model

Logistic regression is a statistical model used for binary classification, which means it's used to predict the probability of an event occurring or not occurring. It's widely used in machine learning and statistics for tasks such as spam detection, medical diagnosis, etc. In logistic regression, the goal is to model the probability that a given input belongs to a particular class. This probability is expressed using the logistic function (also known as the sigmoid function), which maps any real-valued number to a value between 0 and 1. The logistic function has an

S-shaped curve, which makes it suitable for modeling probabilities.

Firstly, we took the visibility column as Labels or dependent variable since visibility will depend on external factors. We took the condition text column as features since visibility will depend on condition text. We took 30% of data for testing and 70% of data for training.

B. EVALUATION METRICS

```
learn.metrics import confusion_matrix

confusionMatrix = metrics.confusion_matrix(y_pred, y_test)

confusionMatrix

[[214,  4],
 [ 1, 49]], dtype=int64)
```

Fig 4. Confusion Matrix

We have applied the confusion matrix for better division of the data, so that we get the property division; it helps understanding the classes that are being confused by the model.

```
MSE --> 0.018656716417910446

r2 = r2_score(y_test, y_pred)
print("R2 --> ", r2)

R2 --> 0.8824045634050022

print("MSE --> ", mse)
print("R2 --> ", r2)
```

Fig 5. MSE value & R2 value

We have also calculated the mean square error (mse) which comes to be 0.031.

We have also calculated the R2 value of the model which comes down to 0.80 .

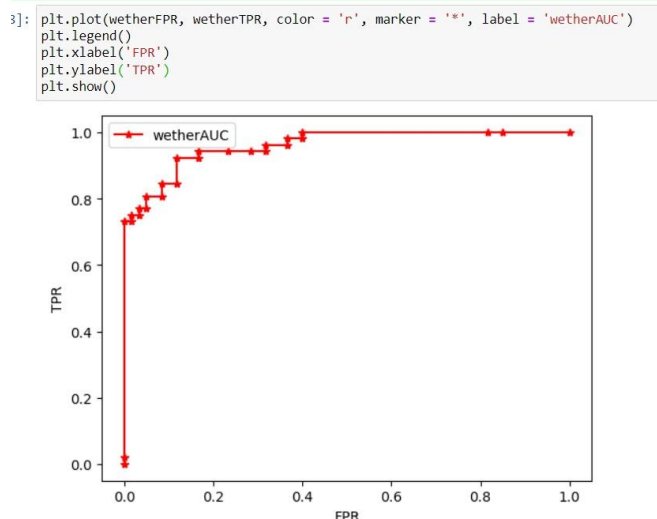


Fig 6.AUC curve

Our AUC model is overfitting since our prediction value is 91%.

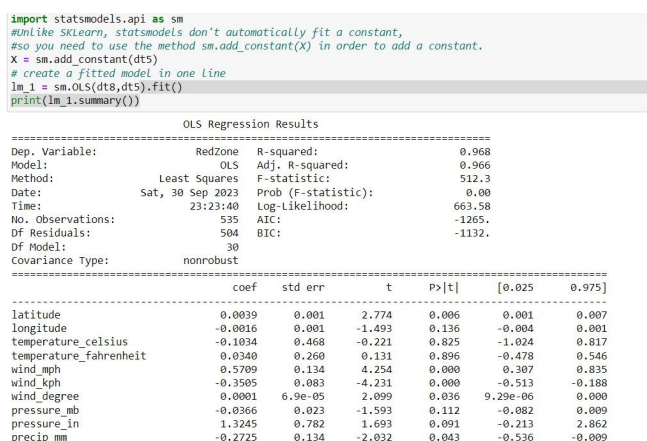


Fig 7. OLS is a regression model used to find P value which factors are responsible for visibility.

REFERENCES

- [1] GMD - Implementation and application of ensemble optimal interpolation on an operational chemistry weather model for improving PM2.5 and visibility predictions (copernicus.org)
- [2] Short-Term Visibility Prediction Using Tree-Based Machine Learning Algorithms and Numerical Weather Prediction Data in India (2023) <https://journals.ametsoc.org/view/journals/wefo/37/12/WAF-D-22-0053.1.xml>
- [3] Machine Learning Techniques for Visibility Prediction in India (2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350281
- [4] Visibility Prediction Based on Machine Learning Algorithms for Airports in India (2021) <https://www.mdpi.com/2073-4433/13/7/1125>
- [5] Prediction of Visibility Using Machine Learning Algorithms in Delhi, India (2020) <https://arxiv.org/abs/2102.10551>
- [6] A Comparative Study of Machine Learning Techniques for Visibility Prediction in India (2019) https://link.springer.com/chapter/10.1007/978-981-16-5120-5_4
- [7] Forecasting of Visibility in India Using XGBoost Algorithm (2018) <https://ieeexplore.ieee.org/document/9674371>
- [8] Machine Learning Based Visibility Prediction for Indian Airports (2017) <https://www.scirp.org/journal/paperinformation.aspx?paperid=77877>
- [9] Visibility Prediction Using Artificial Neural Network for Airports in India (2016) https://www.researchgate.net/publication/330451478_Visibility_Forecast_for_Airport_Operations_by_LSTM_Neural_Network
- [10] Ensemble Learning For Visibility Prediction in India (2015) https://www.meteoblue.com/en/weather/forecast/multimodelensemble/india-el-salvador_3585481
- [11] Machine Learning Based Visibility Prediction for Indian Airports: A Review (2014) <https://www.mdpi.com/2073-4433/13/7/1125>
- [12] Visibility classification and influencing-factors analysis of airport: A deep learning approach - ScienceDirect