***A PROJECT ON***

# “PUNE TEMPERATURE PREDICTION”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYSIS



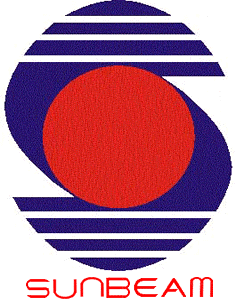
**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

Submitted By:

Pranjali Vanjari (80267)

Aditya Dwivedi (80695)

**Mr.Nitin Kudale Mrs.Manisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘Pune Temperature Prediction’ is done by Pranjali Vanjari & Aditya Dwivedi in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date:

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Pranjali Vanjari

DBDA September 2023 Batch,

SIIT Pune

Aditya Dwivedi

DBDA September 2023 Batch,

SIIT Pune

**TABLE OF CONTENTS**

1. **Introduction**
   1. Introduction And Objectives
   2. Why this problem needs To be Solved?
   3. Dataset Information

## Problem Definition and Algorithm

* 1. Problem Definition
  2. Algorithm Definition

## Experimental Evaluation

* 1. Methodology/Model
  2. Exploratory Data Analysis

## Results And Discussion

1. **GUI**
2. **GitHub link**

## 7.Future Work And Conclusion

* 1. Future Work
  2. Conclusion
     1. **Introduction**
        1. **Introduction And Objectives:**

Weather prediction is a critical task that influences various sectors, including agriculture, transportation, tourism, and disaster management. Accurate weather forecasting can help individuals and organizations make informed decisions, plan activities, and mitigate risks associated with adverse weather conditions.

In this project, we aim to develop a machine learning model to predict weather patterns in Pune, a major city in India. Pune experiences a diverse range of weather phenomena throughout the year, including monsoon rains, hot summers, and cool winters. By leveraging historical weather data and advanced machine learning techniques, we seek to build a model that can forecast various weather parameters such as temperature, humidity, precipitation, and wind speed.

## Why this problem needs To be Solved?

The problem of weather prediction in Pune, and in general, is significant for several reasons:

1. Risk Management: Accurate weather forecasting helps individuals, businesses, and governments mitigate risks associated with adverse weather conditions. It allows farmers to plan crop cultivation and irrigation, helps transportation companies optimize routes and schedules, and enables disaster management agencies to prepare for extreme weather events such as floods, storms, and heatwaves.
2. Resource Allocation: Effective weather prediction aids in the efficient allocation of resources. For example, knowing when and where precipitation is likely to occur can help water resource managers manage reservoirs and water distribution systems more effectively. Similarly, energy companies can optimize electricity generation from renewable sources like solar and wind by forecasting weather patterns accurately.
3. Public Safety: Timely and accurate weather forecasts are crucial for public safety. They enable authorities to issue warnings and advisories to the public, reducing the risk of weather-related accidents and injuries. For instance, advance notice of severe weather events such as hurricanes, tornadoes, or heavy rainfall allows people to take necessary precautions and evacuate vulnerable areas if needed.
4. Economic Impact: Weather affects various sectors of the economy, including agriculture, tourism, construction, and retail. Accurate weather forecasting helps businesses make informed decisions, reduce operational disruptions, and minimize economic losses caused by weather-related events. For instance, retailers can adjust inventory levels based on expected weather conditions to meet consumer demand more effectively.
5. Environmental Sustainability: Weather prediction is essential for environmental monitoring and conservation efforts. It enables scientists to study climate change trends, track the spread of pollutants and contaminants, and assess the impact of human activities on ecosystems. By accurately predicting weather patterns, we can better understand and address environmental challenges such as air and water pollution, habitat destruction, and biodiversity loss.

## Dataset Information.

## Pune.csv

It has 23 columns.

## Problem Definition and Algorithm:

* + - 1. **Problem Definition**

The problem is quite straightforward. Data from Walmart stores accross the US is given, and it is up to us to forecast their weekly sales. The data is already split into a training and a test set, and we want to fit a model to the training data that is able to forecast those weeks sales as accurately as possible. In fact, our metric of interest will be the Mean Absolute Error and R2 score value.The metric is not very complicated. The further away from the actual outcome our forecast is, the harder it will be punished. Optimally, we exactly predict the weekly sales. This of course is highly unlikely, but we must try to get as close as possible.

## Algorithm Definition

**Linear regression:** is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

**Ridge Regression:** Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

**Lasso Regression:** Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.Lasso Regression uses L1 regularization technique,It is used when we have more number of features because it automatically performs feature selection.

**Random forest:** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**Decision Tree:** algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**XGBoost:** or extreme gradient boosting is one of the well-known [gradient](https://analyticsindiamag.com/gradient-descent-everything-you-need-to-know-with-implementation-in-python/) [boosting](https://analyticsindiamag.com/gradient-descent-everything-you-need-to-know-with-implementation-in-python/) techniques (ensemble) having enhanced performance and speed in tree- based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. It is open- source software. Earlier only [python and R packages](https://analyticsindiamag.com/python-vs-scala-for-apache-spark/) were built for XGBoost but now it has extended to Java, Scala, Julia and other languages as well.

## Experimental Evaluation:

* + - 1. **Methodology:**

The objective of this project is to predict the weekly sales of wallmart in US. The data set is contained from Kaggle and has 3 csv files namely features, stores and train. The data is merged to obtain one master datafile and then the data preprocessing is carried out.

## Loading in raw data

features\_df = pd.read\_csv("features.csv") stores\_df = pd.read\_csv("stores.csv") walmart\_df = pd.read\_csv("train.csv")

master\_df =walmart\_df.merge(stores\_df, how='left').merge(features\_df, how='left') print(master\_df.shape)

master\_df.head()

## Preprocessing:

The sales are given for Years 2012-2012 on weekly basis. This data was split to extract information for year, month and week.

master\_df['Date'] = pd.to\_datetime(master\_df['Date'], format='%Y-%m-%d') master\_df['Week\_Number'] = master\_df['Date'].dt.week

master\_df['Month'] = master\_df['Date'].dt.month master\_df["Year"] = master\_df["Date"].dt.year

The data had several missing values and needed to be cleaned. The missing values in ‘Markdown1-5’needed to be cleaned. Since the number of missing values were significant, they were not removed but were replaced with zero.

print(master\_df.isna().sum()) missing\_values = master\_df.isna().sum()

master\_df['MarkDown1'] = master\_df['MarkDown1'].fillna(0) master\_df['MarkDown2'] = master\_df['MarkDown2'].fillna(0) master\_df['MarkDown3'] = master\_df['MarkDown3'].fillna(0) master\_df['MarkDown4'] = master\_df['MarkDown4'].fillna(0) master\_df['MarkDown5'] = master\_df['MarkDown5'].fillna(0) master\_df.isna().sum()

## Flow Diagram :



Start

Data Collection





* + - 1. **Exploratory Data Analysis**

The popularity of each store is plot with the help of a pie chart (fig 2). From the figure we can infer that type A store has the highest popularity followed by type B store and type C store has the least popularity.

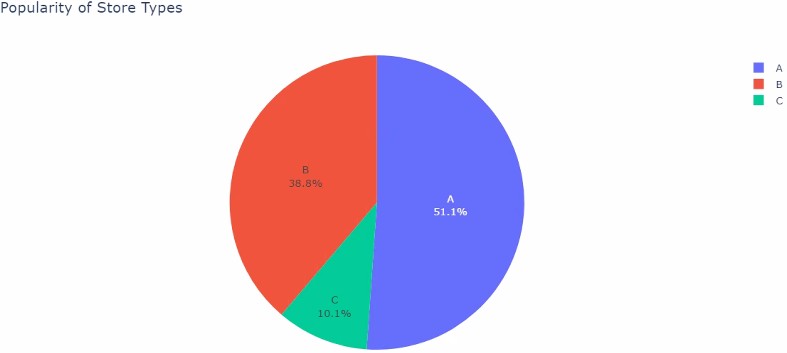


Fig 2: Pie- chart showing store- type wise popularity

The average sale for each store- type is visualized using bar plot (fig 3). From the figure we can infer that type A store has the highest average sales followed by type B store. The type C store has least average sale among the three.

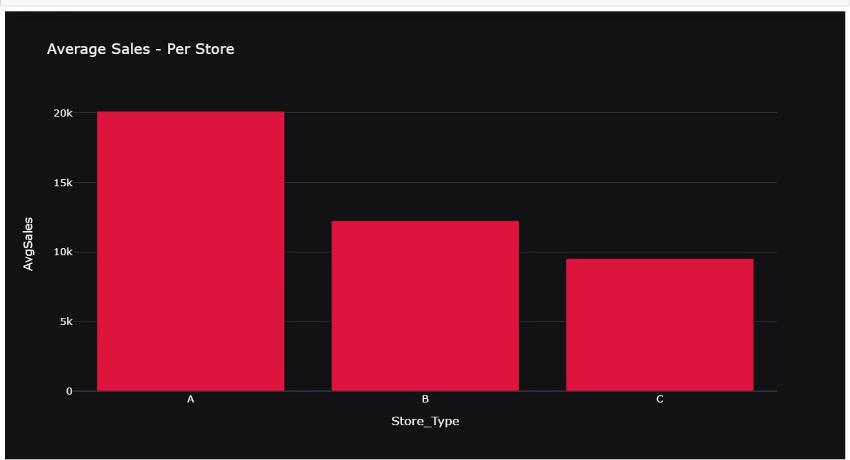


Fig 3: Store- type Vs Average sales

The average sale for each store- type is plotted for each year (fig 4). The plot

shows that Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided

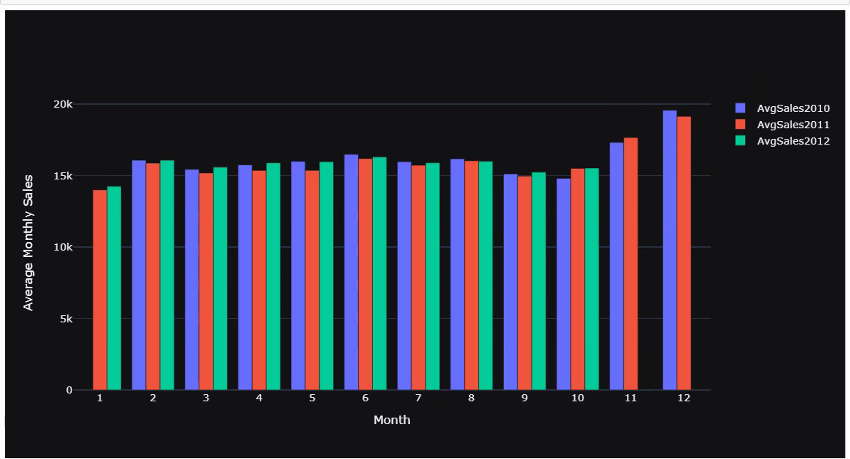


Fig 4: Average monthly sale per year

The average weekly sale per year is plotted using scatter and line plot (fig 5). Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided or any special even.

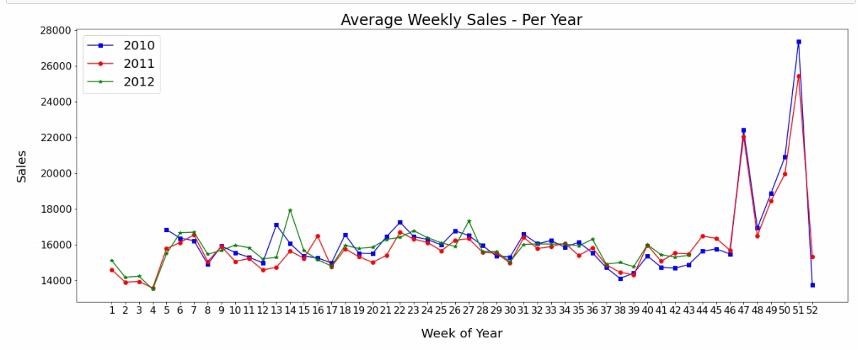


Fig 5: The average weekly sale per year is plotted using scatter and line plot

The average store sales per year is plotted. Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided week of Thanks giving the highest sales in all the 3 years

Average department wise sale per year is plotted for each year. Month of January witnessed the lowest sales for 2011 and 2012 while for 2010 the weekly sales are not given in the data

From Feburary till October the weekly sales nearly remains constant around 15000 for the 3 years November and December showed the highest sales for 2010 and 2011 while for 2012 the sales data has not been provided week of Thanks giving among the 45 stores which have highest average sales

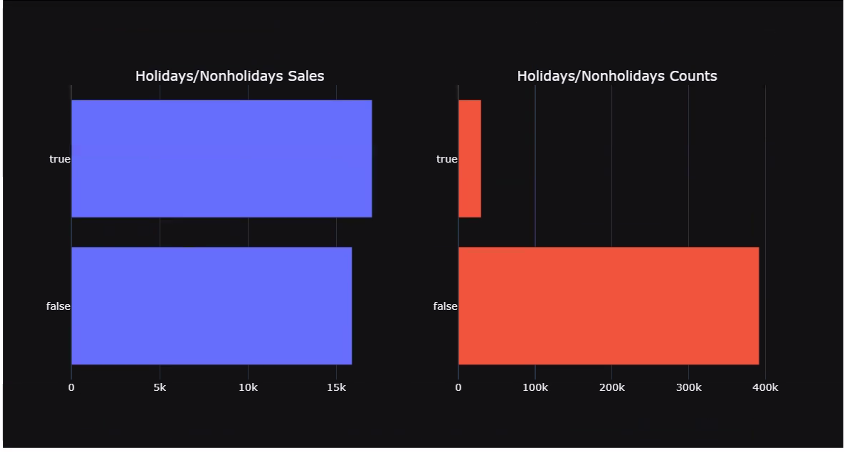


Fig 6: Analysis of sales on holidays and working days

The data is analysed for sales on holidays and other working days. This shows that the sales are comparatively higher on holidays. This information is useful to further improve the store sales. Only 7 percent of the weeks in the data are the holiday weeks Despite being the less peecentage of holiday weeks the sales in the holidays week are on the average higher than in the non-holiday weeks

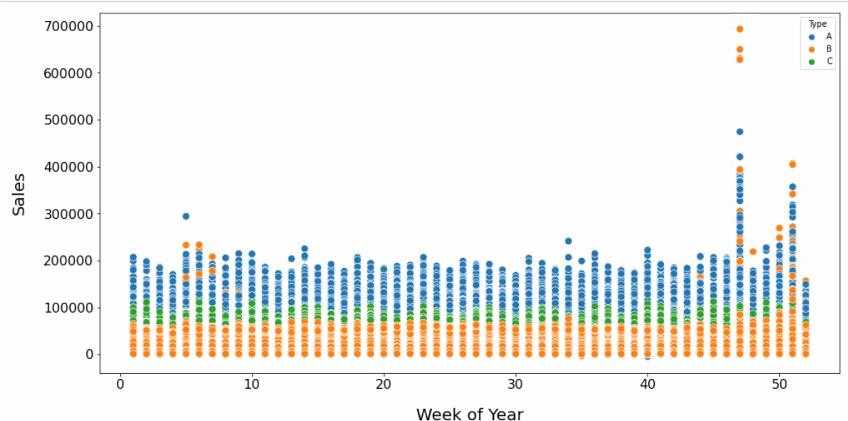


Fig 7: Week of the year vs sales

The sales for each week is plotted for 3 years. This shows only a slight relationship as the weekly sales increased towards the end of the year.

## Results and discussion:

Linear regression, Lasso regression, ridge regression, random forest, decision tree and gradient boosting machine algorithm were used to predict the weekly saes of wallmart. Among the given algorithms Gradient Boosting Machine algorithm was the best performing one as it provided the highest R2 score of 0.94.

from xgboost import XGBRegressor

XGBoost\_model = XGBRegressor() XGBoost\_model.fit(x\_train, y\_train)

y\_prediction = XGBoost\_model.predict(x\_test)

MAE = mean\_absolute\_error(y\_test, y\_prediction) print(f"MAE = {MAE}")

R2 = r2\_score(y\_test, y\_prediction) print(f"R2 = {R2}")

MAE : 1940.99

R2 Score : 0.94

## GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

**6.GitHubLink:**

## 7.Future work And Conclusion 7.1Future Work:

Walmart can analyze the entire store data across US to arrive at an

even more accurate prediction. They can analyze the inventory data as well to optimize their inventory. They can analyze the sales targets and incentives that are given for employees to arrive at achievable sales targets for employees to motivate them better.

## 7.2 Conclusion:

* + - Type 'A' stores are more popular than 'B' and 'C' types
    - Type 'A' stores outclass the 'B' and 'C' types in terms of size and the avergae weekly sales
    - Weekly Sales are effected by the week of year. Holiday weeks witnessed more sales than the non-holiday weeks. Notables are Thanksgiving and Christmas weeks
    - Size of the store is a major contributing factor in the weekly sales
    - Sales are also dependent on the department of the store as different departments showed different levels of weekly sales
    - Among the trained models for predicting the future sales, Gradient Boosting Machine performs the best.