DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES,
NUZVID.

Tourism Demand Forecasting

Report submitted to

Rajiv Gandhi University of Knowledge Technologies,

Nuzvid. for the fulfillment of Mini Project

Of

Bachelor of Technology

in Computer Science and Engineering

by

N180645 (R. Satish)

N180638 (K. Naga Raju)

N180636 (S. Bhargavi)

N181128 (G. Sri Lakshmi)

N180363 (K. Siva Kumar)

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES
DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
RGUKT-Nuzvid, Krishna Dist – 521202

Declaration

We certify that

a. The work contained in this report is original and has been done by us under the guidance of my supervisor(s).

b. The work has not been submitted to any other Institute for any degree or diploma.

c. We have followed the guidelines provided by the Institute in preparing the report.

d. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever We have used materials (data, theoretical analysis, figures, and text) from other sources, We have given due credit to them by citing them in the text of the report and giving their details in the references. Further, We have taken permission from the copyright owners of the sources, whenever necessary.

N180645(R. Satish)

N180638(K. Naga Raju)

N180636(S. Bhargavi)

N181128(G. Sri Lakshmi)

N180363(K. Siva Kumar)

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

RGUKT-Nuzvid, Krishna Dist – 521202

Certificate

This is to certify that the mini project report entitled, "**Tourists Demand Forecasting**" submitted by **Mr. R.Satish, Mr. K.Naga Raju, Ms. S.Bhargavi, Ms. G.Sri Lakshmi, Mr. K.Siva Kumar** to Rajiv Gandhi university of Knowledge Technologies, Nuzvid, India, is a record of bonafide Project work carried out by us under my/our supervision and guidance and is worthy of consideration for the fulfillment of mini-project of Bachelor of Technology in computer Science and Engineering of the Institute.

Mr. Kumar Anurupam Sir	**Examiner**
Project Supervisor Examiner	Project Examiner
Faculty Dept. of CSE	Faculty Dept. of CSE
RGUKT IIIT Nuzvid	**RGUKT IIIT Nuzvid**

ACKNOWLEDGEMENT

We would like to express our profound gratitude and deep regards to our guide ***Mr. Kumar Anurupam*** for his exemplary guidance, monitoring and constant encouragement to us throughout this semester. We shall always cherish the time spent with him during the course of this work due to the invaluable knowledge gained in the field of Machine Learning(ML) and Natural Language Processing (NLP).

We are extremely grateful for the confidence bestowed in us and entrusting our project entitled **"Tourist Demand Forecasting".** We express gratitude to our HOD sir (Dept. of CSE) and other faculty members for being source of inspiration and constant encouragement which helped us in completing the project successfully.

Finally, yet importantly, we would like to express our heartfelt thanks to our beloved **God** and **Parents** for their blessings, our friends for their help and wishes for the successful completion of this project.

N180645(R. Satish)
N180638(K. Naga Raju)
N180636(S. Bhargavi)
N181128(G. Sri Lakshmi)
N180363(K. Siva Kumar)

ABSTRACT

Our tourism demand forecasting model to forecast arrivals for the tourist place Tirumala Tirupati Devasthanam(TTD). We have considered TTD to our model because it maintains historic data of pilgrim arrivals which we access through TTD.News and TTD is popular tourist destination attracting large number of tourists every year. We have considered attributes Day Speciality(weekend or weekday), Weather condition(Based on temperature), Google Trends(frequency of ttd related keywords searched on google), Twitter trend(whether opinion of twitter is +ve,-ve or neutral). Our model uses structured variables to build a tourism demand forecasting model based on Light Gradient Boosting Machine Regressor. LGBMRegresssion uses a leaf-wise tree growth strategy which differs from level-wise strategy employed by many other gradient boosting implementations. In leaf-wise growth, algorithm selects the leaf node with maximum delta loss as the next node to grow. The ensembling algorithm at last forms a improved model. In this approach the google trends, day speciality contributed more information in recognizing the patterns of tourist arrivals. Another three model multiple linear regression, decision tree regression, xgboost(extreme gradient boosting algorithm) also tried to train the model, they have performed considerably but LGBMR gives better results over those three models.

Key Words: Tourism Forecasting, Google Trends, Twitter sentiment analysis, Weather, Day Speciality, Natural Language Processing, Boosting Machine, LGBMR, Machine Learning.

Table Of Contents

List Of Figures

Introduction

All over the world, the tourism industry contributes significantly to economic growth. Thus, forecasting tourist volume is becoming increasingly important for predicting future economic development. Tourism demand forecasting may provide basic information for subsequent planning and policy making. Predicting tourists arrival help destinations to make neccessary arrangements and help tourism businesses to make efficient financial decisions in terms of allocating resouces, staffing, controlling expenditure and exploring oppurtunities for expansion. There was a significant importance of tourism in national economies too.

Tirumala Venkateswara Temple is a Hindu temple in the hill town of Tirumala, near Tirupati in the Chittoor district of Andhra Pradesh. Every year hundreds of thousands of people visits TTD for worshipping Lord Venkateshwara. It is the one the famous tourist attraction place in India, So we considered this place for forecast pilgrims based on keywords. The Tirumala Tirupati Devasthanams (TTD) has proposed Rs 3,096.40 crore budget for 2022-23, which is Rs 158.58 crore more than the revised budget estimates of Rs 2,937.82 crore for 2021-22. As a result, tourism has become ever more important as a driver of Andhra Pradesh state GDP growth. Consequently, this underlines the increasing importance of the tourism sector to the Andhra Pradesh economy.

Therefore, planning marketing strategies for entrepreneurs in the tourism industry is important to attract tourists. Nowadays, tourists also use online media for travel planning, often by searches using a search engine. Potential tourists like to find information before detail planning of their travel. Search engines are thus a part of the ways that tourism operators can know more about tourist interests and anticipate changes in demand, to plan for meeting their expectations. We considered Relevant Internet search keywords cover the various aspects of tourism including Tirupati, Tirumala, VIP darshan, tirumala darshan. Tourism forecasting may depends on other various kinds of factors like tweets which are viral on social media platform like twitter , weather condition (i.e., temperature) and day speciality (i.e., weather it is week day or non-week day).

Machine learning algorithms are used to detect the pattern of trends in tourism arrivals and helps to forecast future arrivals. With the continuous development of the social economy ,the demand for tourist passenger transportation is increasing. The high-intensity and centralized travel demand puts huge pressure on transportation facilities, which may lead to passenger detention, paralysis of local transportation facilities, and even stampede incidents.

Accurate Tourism demand forecasts can provide a reference for the effective allocation of tourism supply resources and help improve the efficiency and safety of tourism travel.

Background and Related Work

In the past, there have been numerous studies related to tourism and the importance of the tourism industry in specific areas. This study focused on forecasting traveler interests using keywords like search engine data, tweets data, weather data, day speciality could predict the inflow of pilgrims.

This study used different regression models to model and forecast no of pilgrims arrivals in a tourist hotspot TTD, based on high correlated keywords like Tirupati, Tirumala, VIP darshan, tirumala darshan in TTD. A variety of machine learning models have been tested for forecasting tourism In most studies the scope is largely determined by the travel area to be studied, and while similar databases are used the differences are in model and purpose.

For example, there are studies on forecasting tourism demand with Google trends, and accuracy comparisons between countries and between cities. The study compared forecasting models based on web search index and/or images of two cities and of two countries. The main objective of the study was to forecast the arrival of tourists in the TTD using data from a search engine such as Google Trends, Historical Data, weather data, Twitter trends. We found that the these features improved forecasting accuracy. In daily tourism volume forecasts were made for tourist attractions. A Light Gradient Boosting Machine Regressor (LGBMR) was proposed for forecasting the no of pilgrims in TTD. There are many variables used in forecasting, namely historical data, search engine data, twitter tweets and weather as independent variables, to forecast the tourism volume of the TTD. Therefore, an LGBMR model is suitable for the task.

Methodology

The proposed approach to forecasting the foreign tourist arrivals, from a search engine by using a LGBMR model presented in Figure1. The tasks mainly fall into some steps, i.e. data collection, data preprocessing, feature selection(determine keywords which have correlation > 0.2 and < -0.2), training phase, validation phase, output prediction.

The first section describes the data collection and preparation. Afterwards, the obtained data will be tested with the Pearson correlation to select the relevant features. The next section involves the modeling to forecast tourist arrivals. Afterwards we will train the model and validate the results. Final section is Output Prediction.

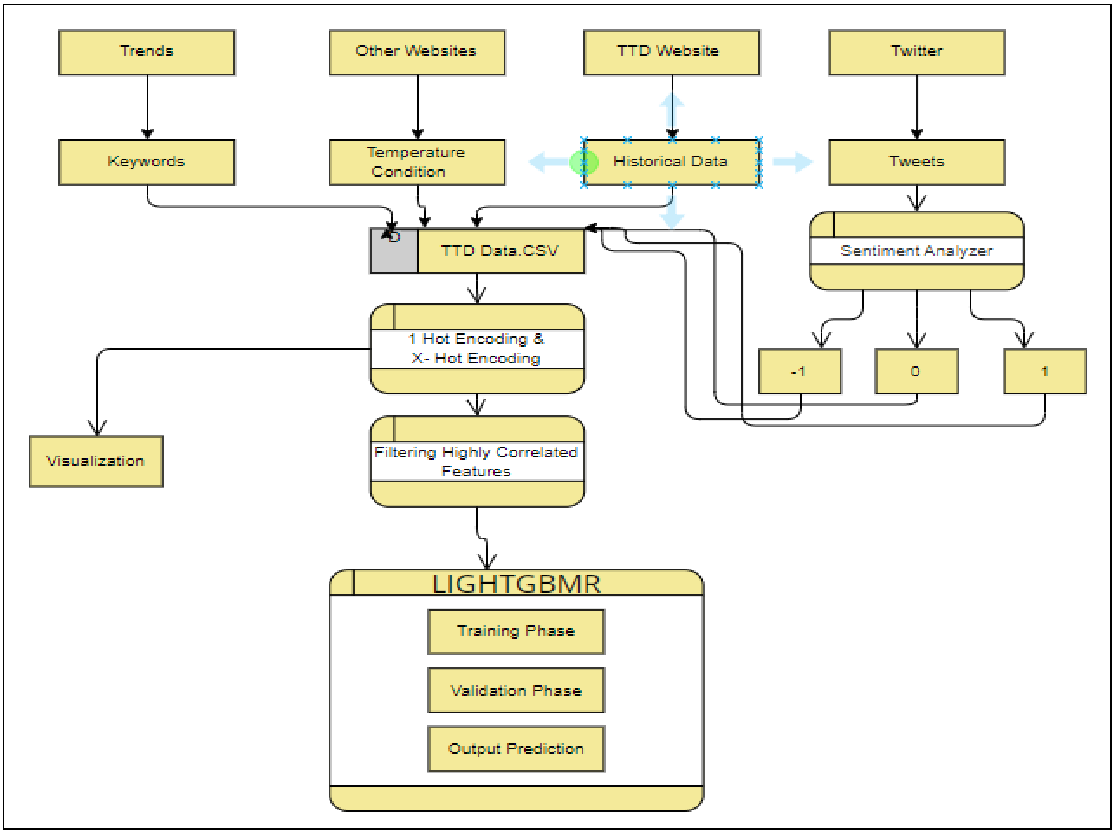

Figure - 1 : The proposed approach to forecasting tourist arrivals

LGBMR Working Architecture:

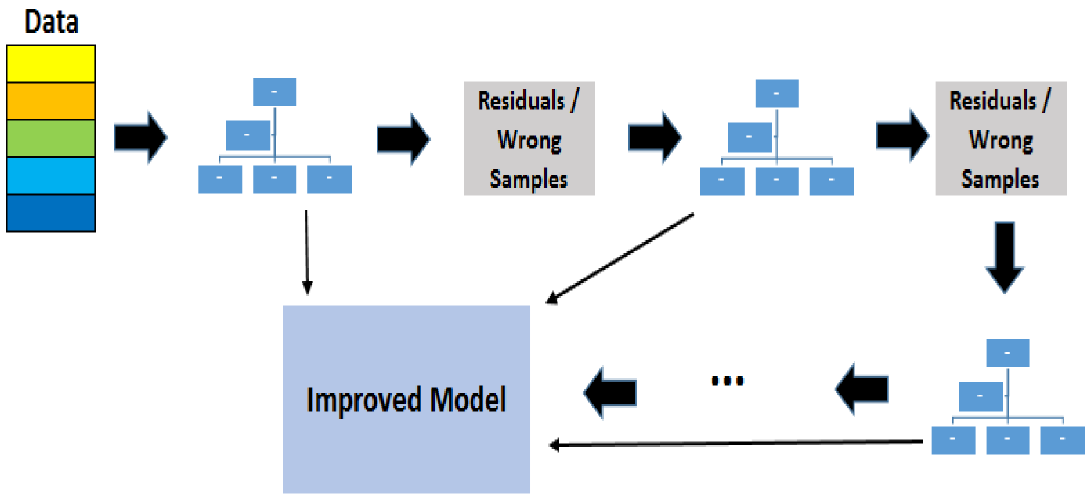

Figure - 2 : LGBMR Model

LightGBM uses a leaf-wise tree growth strategy, which differs from the level-wise strategy employed by many other gradient boosting implementations. In leaf-wise growth, the algorithm selects the leaf node with the maximum delta loss (improvement in the loss function) as the next node to grow. This approach leads to faster convergence and better accuracy.

LightGBM optimizes the gradient descent algorithm by using a technique called Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS selects a small percentage of large-gradient data instances and the full set of small-gradient instances, resulting in a more efficient gradient approximation. EFB combines exclusive feature values to reduce memory usage and speed up training.

Implementation

The implementation of project involved the following steps

4.1 Data Collection:

This step is of utmost importance for the project, as it involves a significant amount of effort to acquire data from various websites and merge them into a unified labeled dataset. It is essential to include an adequate number of instances to effectively train our model to recognize historical trends. After considering the available sources, we have decided to utilize the data from the past year (2022). Specifically, we extracted pilgrim details from the TTD.News website (https://news.tirumala.org/), which provides daily updates on the total number of pilgrims visiting on each day. We also obtained weather data from the Weather Underground website (https://www.wunderground.com/). Additionally, we compiled a list of 15 keywords related to TTD and collected Google Trends data for these keywords, which provides the frequency of searches for a particular keyword on each day. Furthermore, we incorporated the day_speciality values based on the TTD calendar. Another crucial step in data collection involved gathering tweets from Twitter to conduct sentiment analysis on social networking data. To ensure the reliability of the collected tweets, we only considered those with a minimum of 5 likes and 2 retweets. Since all the data was collected based on specific dates, we then consolidated the entire dataset into a single entity according to the dates.

It's important to note that the data extraction from each website was performed using the Selenium WebDriver in Python.

Date	Pilgrims	Tonsures	Hundi	Special Day	Temperature	Condition
01/01/2022	36,560	14,084	Rs. 2.15 Cr	New Years Day	+77°	Moderate rain
02/01/2022	38,894	12,270	Rs. 3.93 Cr	Amavasya, Sravanam, Tirupati Sri G.T Adhyay	+79°	Few clouds
03/01/2022	31,776	16,046	Rs. 2.69 Cr	Maha Sivaratri, Tirupati Sri KT Nandi Vahanar	+79°	Few clouds
04/01/2022	31,523	14,692	Rs. 2.45 Cr	Amavasya	+79°	Few clouds
05/01/2022	32,044	17,558	Rs. 2.61 Cr	May Day	+79°	Few clouds
06/01/2022	29,652	14,916	Rs. 2.75 Cr	Buddha Jayanthi, Karvetinagaram Sri VGS Pus	+81°	Few clouds
07/01/2022	62,856	22,115		normal	+81°	Partly cloudy
08/01/2022	33,619	15,769	Rs. 2.42 Cr	Naga Chaturthi	+81°	Partly cloudy
09/01/2022	32,235	15,003	Rs. 3.19 Cr	Rushi Panchami	+84°	Partly cloudy
10/01/2022	32,242	15,715	Rs. 2.71 C	Tirumala Sri TT Garuda Sava	+82°	Few clouds
11/01/2022	23,744	12,017	Rs. 2.50 Cr	Sravanam, Tirumala Sri TT Pushpa Yagam	+82°	Broken clouds
12/01/2022	25,524	13,052	Rs. 1.59 Cr	normal	+84°	Few clouds
13/01/2022	46,118	10,594	Rs. 4.09 Cr	Vaikunta Ekadasi, Sri T.T Vaikunta Dwara Dar	+79°	Heavy rain in places
14/01/2022	37,304	9,645	Rs. 2.13 Cr	Bhogi	+82°	Moderate rain
15/01/2022	34,375	11,156	Rs. 2.92 Cr	Makara Sankranthi, Sanitrayodashi	+81°	Few clouds
16/01/2022	35,642	11,178	Rs. 2.77 Cr	Kanuma, Sri Godadevi Parinayotsavam	+79°	Few clouds
17/01/2022	35,333	12,252	Rs. 2.52 Cr	Purnima, Punarvasu, Sri Ramakrishna Teerth	+79°	Broken clouds
18/01/2022	33,971	11,356	Rs. 2.62 Cr	normal	+79°	Light rain
19/01/2022	34,187	13,279	Rs. 2.11 Cr	normal	+82°	Clear sky
20/01/2022	36,092	13,738	Rs. 2.58 Cr	normal	+84°	Clear sky
21/01/2022	39,440	13,692	Rs. 2.53 Cr	normal	+86°	Partly cloudy
22/01/2022	45,481	15,909	Rs. 2.33 Cr	Sri T.T Vaikunta Dwara Darshanam Ends	+88°	Few clouds
23/01/2022	27,895	13,631	Rs. 3.48 Cr	normal	+86°	Few clouds
24/01/2022	27,223	14,624	Rs. 1.87 Cr	normal	+86°	Broken clouds

Figure 3 : Unpreprocessed Historical Data

Date	Tweets
Jan 1, 2022	[]
Jan 2, 2022	[]
Jan 3, 2022	['తిరుమల:శ్రీవారి దర్శనానికి సంబంధించి నకిలీ టికెట్లు విక్రయిస్తున్న దళారులపై విజిలెన్స్ అధికారుల ఫిర్యాదులు. రూ. 300 ప్రత్యేక ప్రవేశ దర్శనం నకిలీ టికెట్లు రూ.3300 నుంచి 7వేలకు విక్రయించిన திருமலை திருப்பதி தேவஸ்தானத்தில் அரசியல் தலையீட்டால் நடைபெறும் மிகப் பெரிய ஊழல் மிக விரைவில் அம்பலமாகும் என எதிர்பார்க்கலாம்.', 'Prepare detailed
Jan 4, 2022	[]
Jan 5, 2022	['', '\n3,402 ఎకరాల భూమికి సంబంధించి 24 ఏళ్లుగా జరుగుతున్న పోరాటంలో టీటీడీ విజయం ', '\n✓నిన్న తిరుమల శ్రీవారిని 31,523 మంది భక్తులు దర్శించుకున్నారు, తలనీలాలు సమర్పించిన 14,69
Jan 6, 2022	[]
Jan 7, 2022	['\nతిరుమల: బాలాజీ ఆరోగ్య వరప్రసాదం పథకానికి కోటి రూపాయలు విరాళంగా అందించిన మహారాష్ట్రకి చెందిన భక్తులు', '\n✓నిన్న తిరుమల శ్రీవారిని 32,613 మంది భక్తులు దర్శించుకున్నారు, తలనీ
Jan 8, 2022	['\nప\u200cర్వాల కార\u200cణంగా ద\u200cర్శ\u200cనం టికెట్లు క\u200cలిగి ద\u200cర్శ\u200cనం చేసుకోలేక\u200cపోయిన భ\u200cక్తులకు ఆరు నెల\u200cల్లోపు స్వామివారి ద\u200cర్శ\u200cనం చేస
Jan 9, 2022	[]
Jan 10, 2022	[]
Jan 11, 2022	['', '\nVehicles allowed to ply on the ghat road leading to ', '\nతిరుమల: ఇవాళ రాత్రి నుంచి అందుబాటులోకి రానున్న రెండో ఘాట్\u200c రోడ్డు.. ఎల్లుండి నుంచి 10 రోజుల పాటు శ్రీవారి ఆలయంలో వైకు
Jan 12, 2022	[]
Jan 13, 2022	['\nవైకుంఠ ఏకాదశి పర్వదినాన్న తిరుమల లో భక్తుల ఆగ్రహం. అర్ధరాత్రి నుండి క్యూ లైన్ లో వున్నా తితిదే అధికారులు తమను సరిగ్గా పట్టించుకోవడం లేదని తీవ్ర ఆగ్రహం. తితిదే చైర్మన్, రాష్ట్ర ప్రభుత
Jan 14, 2022	["\nPeople cant even get a proper Darshana in TTD Temple and can't even express his views on "]
Jan 15, 2022	[]
Jan 16, 2022	['', '\nPeople cant even get a proper Darshana in ']
Jan 17, 2022	['', '\nSince ']
Jan 18, 2022	["\nVaikunta Ekadasi spreading devotional vibes with this beautiful lighting 'Tirumala' \n\nCredits: \n@radha127\n "]
Jan 19, 2022	['\n✓నిన్న తిరుమల శ్రీవారిని 33,971 మంది భక్తులు దర్శించుకున్నారు, తలనీలాలు సమర్పించిన 11,356 మంది భక్తులు, హుండీ ఆదాయం రూ. 2.62 కోట్లు\n ']
Jan 20, 2022	[]
Jan 21, 2022	['', '\nAkkineni Nagarjuna & Amala at Tirumala \n\nశ్రీవారిని దర్శించుకున్న అక్కినేని నాగార్జున దంపతులు\n\nప్రజలందరూ బాగుండాలని స్వామివారిని కోరుకున్నాను.']
Jan 22, 2022	['\n"Sri Venkateswara Central ', '\n✓తిరుమ\u200cల: నేటితో ముగియ\u200cనున్న శ్రీ\u200cవారి వైకుంఠ ద్వారా ద\u200cర్శ\u200cనాలు.. 10 రోజుల పాటు కొన\u200cసాగిన వైకుంఠ ద్వార ద\u200cర్శ\u

Figure 4 : Unpreprocessed Tweets Data

4.2 Feature Engineering(Data Preprocessing):

Similar to any machine learning project, a significant portion of the project timeline is dedicated to data preprocessing. In our specific case, the data we are dealing with is real-time and encompasses various unstructured formats, making the process of transforming the dataset into a clean and structured format quite labor-intensive.

To address missing values in the pilgrim data, we employed a strategy of substituting empty values with the mean value of corresponding weekdays. As for the Google Trends data related to keywords, the format remained consistent throughout the preprocessing phase. Regarding the weather data, we converted the temperature values from the Fahrenheit (°F) scale to a numerical format by removing the associated symbol.

When more than one category is active in a one-hot encoded representation, it is commonly referred to as "multi-hot encoding" or "multi-label encoding". In this encoding scheme, multiple categories can be simultaneously active for a given observation, and the binary vector representation would have multiple elements set to 1. The day speciality feature is then done multi-hot encoding with elements as Festivals, Tidi, Seva_Begins, Sevas, Public_Day, Normal with more than one element may active at a time

The above procedures were necessary to ensure that the dataset was appropriately formatted and ready for subsequent analysis and modeling. Such data preprocessing efforts are crucial in order to maximize the accuracy and reliability of the machine learning algorithms applied to the dataset.

Natural Language Preprocessing:

The tweets extracted from Twitter encompass multiple languages, requiring us to unify them into a single language for sentiment analysis. To achieve this, we utilized the Google Translator to convert the tweets into plain English, ensuring that the information remains consistent across all tweets.

Next, in order to process the tweets using machine learning algorithms, we employed our own sentiment analyzer as well as Python's built-in sentiment analyzer. Comparatively, the built-in sentiment analyzer yielded superior results. Therefore, each tweet was passed through the sentiment analyzer, assigning a label based on the score obtained. Tweets with a score below -0.4 were labeled as -1, representing a negative trend, while scores between -0.4 and 0.4 were labeled as 0, indicating a neutral trend. Scores above 0.4 were labeled as 1, reflecting a positive trend. It is important to note that days without any tweets were considered as having a neutral trend.

Ultimately, we derived a "Review" column containing values (-1, 0, 1), which indicates the sentiment expressed on Twitter for each corresponding day.

Date	Piligrims	Tonsures	Hundi	Temperat	Condition	ttd online	Tirupati	Tirumala	VIP darsha	How to re	ttd sarva d	ttd free da	apsrtc tiru	apsrtc ttd	ksrtc tirup	srinivasam	tirumala d
01/01/2022	36560	14084	2.15	77	1275	0	43	23	16	0	0	0	32	0	0	0	0
02/01/2022	38894	12270	3.93	79	972	6	41	25	14	0	0	0	0	0	0	0	0
03/01/2022	31776	16046	2.69	79	972	10	40	19	0	0	0	14	24	0	0	0	0
04/01/2022	31523	14692	2.45	79	972	0	45	17	22	0	24	14	0	0	53	0	0
05/01/2022	32044	17558	2.61	79	972	8	32	14	21	0	30	10	0	0	0	0	18
06/01/2022	29652	14916	2.75	81	972	0	33	13	0	0	16	0	0	0	31	25	11
07/01/2022	62856	22115	2.21	81	1324	10	40	19	0	0	0	0	0	0	0	0	0
08/01/2022	33619	15769	2.42	81	1324	0	36	20	11	0	0	0	0	0	36	0	0
09/01/2022	32235	15003	3.19	84	1324	0	33	15	0	0	0	11	0	30	0	21	0
10/01/2022	32242	15715	2.71	82	972	0	35	15	0	31	0	11	0	54	0	0	0
11/01/2022	23744	12017	2.5	82	1291	0	32	16	32	0	0	0	0	63	0	25	9
12/01/2022	25524	13052	1.59	84	972	9	34	19	0	0	34	0	0	0	0	0	9
13/01/2022	46118	10594	4.09	79	1878	7	30	18	17	0	0	10	0	0	50	0	14
14/01/2022	37304	9645	2.13	82	1275	0	25	14	0	0	0	34	0	0	0	44	11
15/01/2022	34375	11156	2.92	81	972	0	30	15	0	0	0	0	0	0	0	41	32
16/01/2022	35642	11178	2.77	79	972	0	27	16	20	0	0	0	0	0	24	47	0
17/01/2022	35333	12252	2.52	79	1291	0	29	13	0	0	0	9	18	0	37	0	21
18/01/2022	33971	11356	2.62	79	962	9	30	13	34	0	0	10	0	0	0	0	0
19/01/2022	34187	13279	2.11	82	862	0	29	16	0	0	0	0	0	0	25	0	11
20/01/2022	36092	13738	2.58	84	862	15	32	18	0	0	0	10	0	0	0	23	33
21/01/2022	39440	13692	2.53	86	1324	11	34	21	0	91	0	16	39	0	0	0	0

Figure – 5 : Preprocessed Data

apsrtc tiru	apsrtc ttd	ksrtc tirup	srinivasam	tirumala d	tirumala d	tirupati ba	tirupati da	ttd free da	Week_day	Review	Timestamp	Day	Speciality	Festivals	Tidi	Seva_Begi	Sevas	Public_Da	Normal
32	0	0	0	0	9	0	0	0	1	-1	1640975400	Saturday	New Year	0	0	0	0	1	0
0	0	0	0	0	9	0	0	0	1	0	1641061800	Sunday	Amavasya	0	1	1	0	0	0
24	0	0	0	0	5	7	0	14	0	0	1641148200	Monday	Maha Siva	1	0	0	1	0	0
0	0	53	0	0	16	0	0	14	0	0	1641234600	Tuesday	Amavasya	0	1	0	0	0	0
0	0	0	0	18	10	7	0	10	0	1	1641321000	Wednesd	May Day	0	0	0	0	1	0
0	0	31	25	11	3	8	25	0	0	0	1641407400	Thursday	Buddha Ja	0	0	0	1	1	0
0	0	0	0	0	7	7	0	0	0	0	1641493800	Friday	normal	0	0	0	0	0	1
0	0	36	0	0	11	0	0	0	1	-1	1641580200	Saturday	Naga Chat	0	1	0	0	0	0
0	30	0	21	0	9	0	0	11	1	0	1641666600	Sunday	Rushi Pan	0	1	0	0	0	0
0	54	0	0	0	8	8	0	11	0	0	1641753000	Monday	Tirumala S	0	0	0	1	0	0
0	63	0	25	9	7	13	0	0	0	0	1641839400	Tuesday	Sravanam	0	1	0	1	0	0
0	0	0	0	9	9	0	0	0	0	0	1641925800	Wednesd	normal	0	0	0	0	0	1
0	0	50	0	14	11	0	43	10	0	-1	1642012200	Thursday	Vaikunta	0	1	1	0	0	1
0	0	0	44	11	6	12	0	34	0	0	1642098600	Friday	Bhogi	1	0	0	0	0	0
0	0	0	41	32	11	0	0	0	1	0	1642185000	Saturday	Makara Sa	1	1	0	0	0	0
0	0	24	47	0	3	0	69	0	1	0	1642271400	Sunday	Kanuma, S	1	0	0	1	0	0
18	0	37	0	21	8	0	0	9	0	0	1642357800	Monday	Purnima,	0	1	0	1	0	0
0	0	0	0	0	2	0	0	10	0	1	1642444200	Tuesday	normal	0	0	0	0	0	1
0	0	25	0	11	7	12	0	0	0	1	1642530600	Wednesd	normal	0	0	0	0	0	1
0	0	0	23	33	12	0	67	10	0	0	1642617000	Thursday	normal	0	0	0	0	0	1
39	0	0	0	0	12	0	0	16	0	0	1642703400	Friday	normal	0	0	0	0	0	1
0	61	0	0	28	12	0	60	16	1	0	1642789800	Saturday	Sri T.T Vai	0	0	0	1	0	0

Figure – 6 : Preprocessed Data

Nov 27, 2022	['The congestion of devotees who continue in Tirumala. 22 devotees waiting in compartments. 40 hours for devotees who do not have tokens. 76,681 devotees who visited Srivari yesterday. 40,109 devotees who submitted headlines. Yesterday Srivari Hundi's income was Rs.3.38 crores.
Nov 28, 2022	['Panchami Theertham at Padmavathi Ammavari Temple, Tiruchanur', 'Padmavati Amman's Brahmotsavas that reached the last event. Today Padmavati Amman Panchami Theertham. Yellow saffron sare for the seller from Thirumala Srivari. At 11.40 am, the seller. Tirupati roads with devotees. ']
Nov 29, 2022	['Thirumala: Tomorrow TTD Governing Council Meeting']
Nov 30, 2022	[]
Dec 1, 2022	['VIP Break Changes .. Experimental Success! Common devotees are our priority: TTD Chairman YV Subbaradi Garu', 'VIP Break Changes in VIP Break Darshana .. Experimental Successmanya Bhaktu Gold tasks. Thirumala Tirupati Temple is making arrangements. The governing body has decided to start work from March. The Hindu Dharma campaign is to be handed over to the Department of Dept. and District Collectors. The decision should be withdrawn immediately and the program should be carried out by the Samaritan organization. ',' Ananda Nilayam Golden February 23 will begin on the 23rd .. In 6 months, the Brahmotsava gifts for the full regular, contract and outsourcing employees have been revealed by TTD Chairman Subbara
Dec 2, 2022	[]
Dec 3, 2022	[]
Dec 4, 2022	[',' TTD Board Number ... Literary Real Estate chief Lakshi Narayana arrested .... 4500 families of road milk ... 530 crores of real estate ventures for gharana .... cheating ..... · " And 'Satyameva Jayate has collected Rs 2,500 to Rs 900 crore and is a criminal ritual, such as Sahiti Lakshmi Narayana, who have committed fraud. 'Most precious seconds' and' Tirumala rushed devotees who grew up in Thirumalai yesterday, 69,931 devotees presented to the Swami 34,813 Hundi Income
Dec 5, 2022	[We Went Crazy In The Wild Card ... Stroom for Karnataka devotees in tirupati is under construction in Tirupati is in Tirupati, 'Thirumala, 80,001 devotees who have visited the rush of devotees in Tirumala, 32,967 devotees who have been presented to t 32,967 devotees and Hundi income of Rs.

Figure – 7 : Translated Tweets Data

4.3 Feature Selection:

With the labelled-dataset, there may redundant or unwanted features which will effect the model performance and computation. Feature selection is the process of selecting a subset of relevant features from a larger set of available features to improve the performance and efficiency of a machine learning model. The goal is to identify and retain the most informative features while discarding redundant or irrelevant ones.

To get relevant variables, we consider only those features who correlation with output(i.e pilgrims) is < -0.2 and > 0.2. Then only these features are given to algorithm for model construction.

```
Variables                           Correlation_co-efficient
===========                         ========================
('Pilgrims', 'Pilgrims')            0.9999999999999982
('Pilgrims', 'Tonsures')            0.8572929983934131
('Pilgrims', 'Hundi')               0.6399828942919743
('Pilgrims', 'Temperature')         0.2724375735704814
('Pilgrims', 'Tirupati')            0.5949582772666608
('Pilgrims', 'Tirumala')            0.332022380013128
('Pilgrims', 'VIP darshan')         0.3088058993645298
('Pilgrims', 'Week_day')            0.24441067668084554
('Pilgrims', 'Timestamp')           0.516230692026717
```

Figure – 8 : Correlation Data

4.4 Traning Phase:

The training phase of LightGBMRegression (LGBMR) involves the iterative process of building an ensemble of decision trees to minimize the specified loss function and optimize the model's predictive performance.

The LightGBMRegression model is fitted with the training dataset and it will form a tree by repeating the boosting iterations and constructing multiple decision trees.

4.5 Validation phase:

The validation phase in LightGBM Regression (LGBMR) involves assessing the performance of the trained model on a separate validation dataset. This phase helps evaluate how well the model generalizes to unseen data and provides insights into its predictive capabilities.

The predicted values from the LGBMR model on the validation dataset are compared with the corresponding true values of the target variable. This comparison is used to evaluate the model's predictive performance and assess its accuracy.

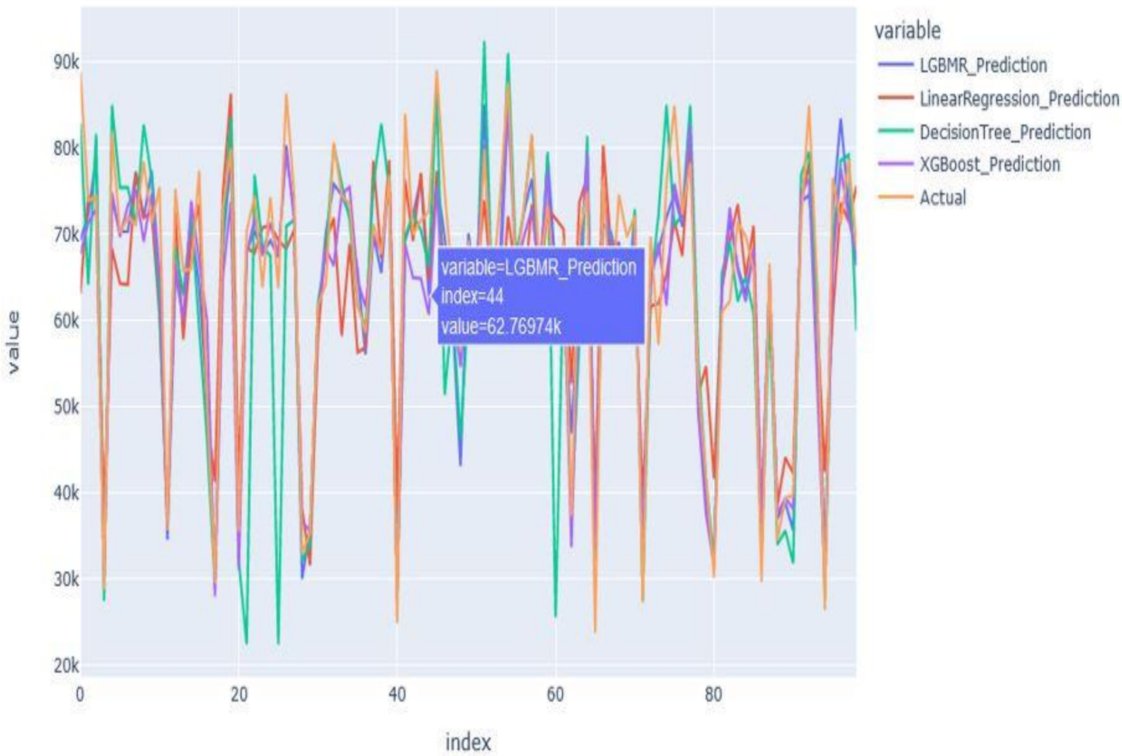

Figure - 9 : Validation Phase

4.6 Web Application Development

The interface of the application asks the user to enter the attributes which are taken by the model as input i.e Weather day is week_day or not, Temperature of the day, Related google trends, Reviews are positive or negative and it redirects to the output page predicts the possible number of visitors.

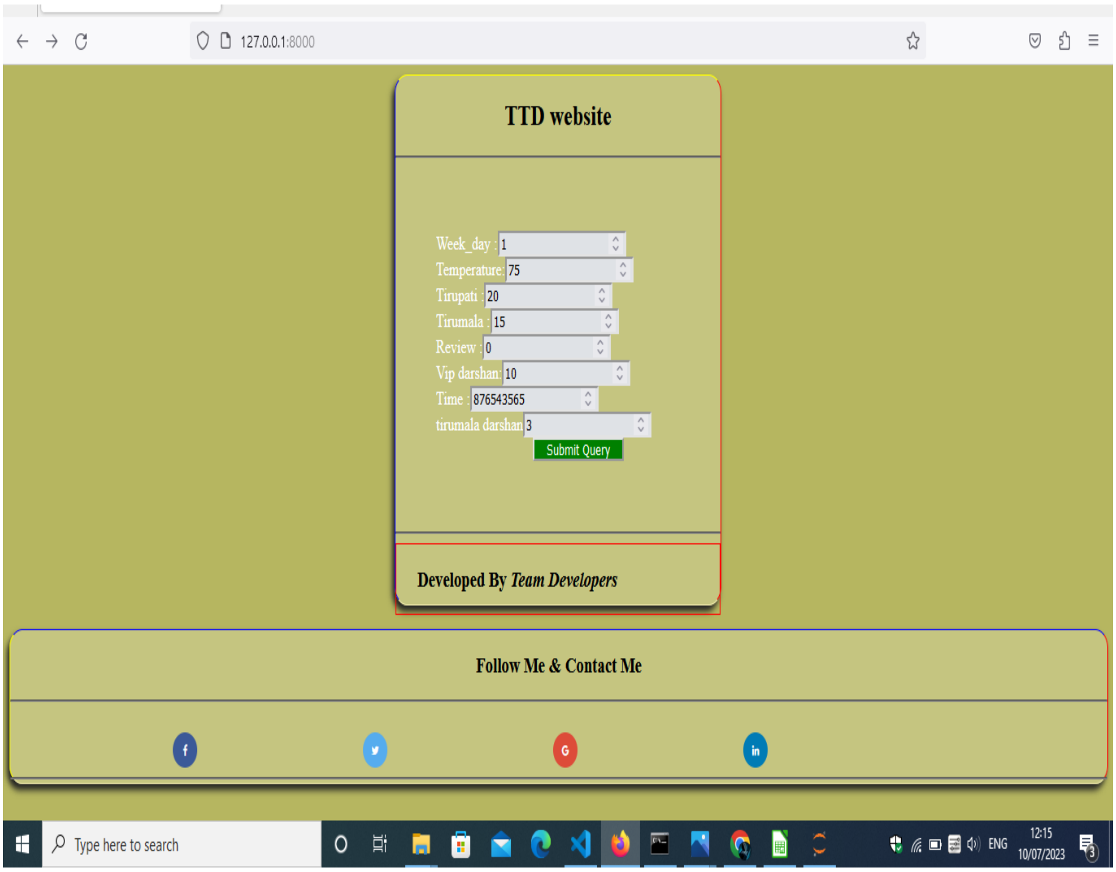

Figure – 10 : Web Interface

4.7 Output Prediction

In LightGBM Regression (LGBMR), the output prediction refers to the process of using the trained model to make predictions on new, unseen data. Once the model has been trained and validated, it can be applied to new instances to estimate their corresponding target variable values.

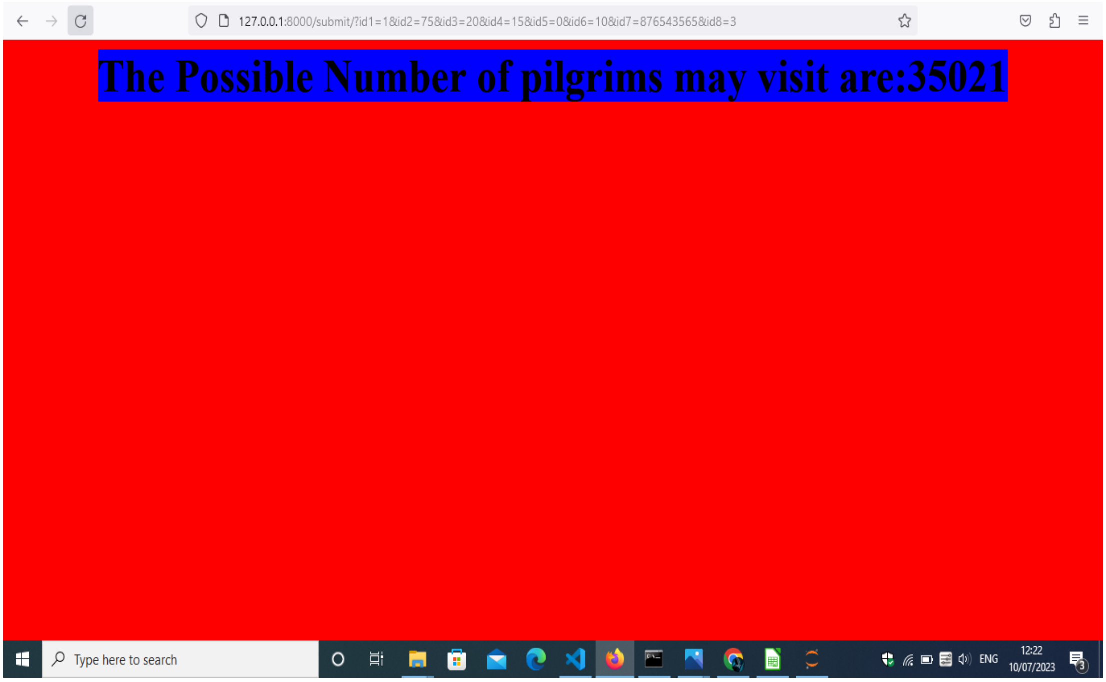

Figure – 11 : Sample Output Data

Dataset Description

The dataset used in this project is a comprehensive collection of data specifically curated for the purpose of predicting tourism demand. It was meticulously gathered through web scraping techniques, employing the powerful Selenium framework for automated data extraction from various online sources.

The dataset encompasses a variety of features, carefully selected to capture multiple factors that influence tourism demand. These features include:

Historic Data: Historical records related to tourism demand, such as past visitor arrivals,Hundi kanuka`s and tonsures. This allows for an analysis of trends and patterns over time.

Google Trends: Data obtained from Google Trends, specifically capturing the search index intensity related to Thirumala Thirupathi Devasthanam (TTD) keywords. By collecting this data, we gain insights into the popularity and search interest surrounding TTD as a tourist destination..

Tweets: Data sourced from Twitter, specifically capturing tweets related to Thirumala Thirupathi Devasthanam (TTD) like #TTD. By collecting tweets specifically related to TTD, insights can be gained into the sentiment, preferences, discussions, and opinions surrounding TTD as a tourist destination.

Weather Conditions: Information about weather conditions, such as temperature. Weather plays a crucial role in influencing tourist behavior and destination choices.

Weekday: The day of the week (e.g., Monday, Tuesday) when the data was collected. This feature allows for the consideration of any variations in tourism demand based on specific weekdays.Later we converted this weekday into binary attribute like holiday or not.That makes valuable changes in the prediction.Our accuracy have been increased.

Each feature within the dataset is structured and organized, allowing for efficient data analysis and modeling. Data preprocessing techniques were applied to handle missing values, address outliers, and standardize the data, ensuring the dataset's quality and suitability for machine learning algorithms.

The carefully curated dataset, incorporating historic data, Google Trends, tweets, weather conditions, and weekday information, serves as the foundation for training and evaluating the tourism demand prediction model. By incorporating these diverse features, the dataset captures the various aspects that impact tourism demand, providing valuable insights for accurate forecasting.

It's important to note that the dataset is continually updated and can be expanded with additional relevant data sources. This allows for ongoing refinement of the model and adaptation to changing patterns and trends in tourism demand

Date	Piligrims	Tonsures	Hundi	Temperat	Condition	ttd online	Tirupati	Tirumala	VIP darsha	How to re	ttd sarva d	ttd free da	apsrtc tiru	apsrtc ttd	ksrtc tirup	srinivasan	tirumala d	tirumala (
01/01/2022	36560	14084	2.15	77	1275	0	43	23	16	0	0	0	32	0	0	0	0	9
02/01/2022	38894	12270	3.93	79	972	6	41	25	14	0	0	0	0	0	0	0	0	9
03/01/2022	31776	16046	2.69	79	972	10	40	19	0	0	0	14	24	0	0	0	0	5
04/01/2022	31523	14692	2.45	79	972	0	45	17	22	0	24	14	0	0	53	0	0	16
05/01/2022	32044	17558	2.61	79	972	8	32	14	21	0	30	10	0	0	0	0	18	10
06/01/2022	29652	14916	2.75	81	972	0	33	13	0	0	16	0	0	0	31	25	11	3
07/01/2022	62856	22115	2.21	81	1324	10	40	19	0	0	0	0	0	0	0	0	0	7
08/01/2022	33619	15769	2.42	81	1324	0	36	20	11	0	0	0	0	0	36	0	0	11
09/01/2022	32235	15003	3.19	84	1324	0	33	15	0	0	0	11	0	30	0	21	0	9
10/01/2022	32242	15715	2.71	82	972	0	35	15	0	31	0	11	0	54	0	0	0	8
11/01/2022	23744	12017	2.5	82	1291	0	32	16	32	0	0	0	0	63	0	25	9	7
12/01/2022	25524	13052	1.59	84	972	9	34	19	0	0	34	0	0	0	0	0	9	9
13/01/2022	46118	10594	4.09	79	1878	7	30	18	17	0	0	10	0	0	50	0	14	11
14/01/2022	37304	9645	2.13	82	1275	0	25	14	0	0	0	34	0	0	0	44	11	6
15/01/2022	34375	11156	2.92	81	972	0	30	15	0	0	0	0	0	0	0	41	32	11
16/01/2022	35642	11178	2.77	79	972	0	27	16	20	0	0	0	0	0	24	47	0	3
17/01/2022	35333	12252	2.52	79	1291	0	29	13	0	0	0	9	18	0	37	0	21	8
18/01/2022	33971	11356	2.62	79	962	9	30	13	34	0	0	10	0	0	0	0	0	2
19/01/2022	34187	13279	2.11	82	862	0	29	16	0	0	0	0	0	0	25	0	11	7
20/01/2022	36092	13738	2.58	84	862	15	32	18	0	0	0	10	0	0	0	23	33	12
21/01/2022	39440	13692	2.53	86	1324	11	34	21	0	91	0	16	39	0	0	0	0	12

Figure – 12 : Data Set

Results

In this section, we present the results of our tourist demand prediction model for Thirumala Thirupathi Devasthanam (TTD). The dataset used for training and evaluation included features such as Twitter tweets, Google Trends data, historic data, day type, and weather conditions.

We employed four different machine learning algorithms - LightGBM, Decision Tree, Regression, and XGBoost - to predict tourist demand for TTD. Each algorithm was trained and evaluated using the collected dataset.

To assess the performance of the algorithms, we used various evaluation metrics, including mean absolute error (MAE), root mean squared error (RMSE), and R-squared value (R^2). These metrics allow us to measure the accuracy and predictive power of our models.

The results of our analysis revealed promising performance from all four algorithms in predicting tourist demand for TTD. LightGBM exhibited the lowest MAE and RMSE values, indicating superior accuracy compared to the other algorithms. Furthermore, LightGBM attained the highest R^2 value, signifying a strong correlation between the predicted and actual tourist demand.

To visually demonstrate the results, we utilized Plotly, a powerful data visualization library. We created interactive graphs and charts to showcase the performance of the different algorithms and compare their predictions with the actual tourist demand.

Overall, the results of our tourist demand prediction model highlight its potential usefulness in forecasting future tourist demand for Tirumala Tirupathi Devasthanam (TTD). However, further evaluation and refinement may be necessary to enhance the accuracy and robustness of the model.

These findings can assist TTD in making informed decisions related to resource allocation, crowd management, and planning for visitor services. By accurately predicting tourist demand, TTD can optimize its operations and provide an enhanced experience for visitors.

Model	R^2 Score	Mean Squared Error	Mean Absolute Error
LightGradientBoostingMachine	0.862922	37,141,157.632963	4,772.399706
LinearRegression	0.762544	64,338,423.038450	6,389.992184
DecisionTreeRegression	0.651884	94,321,403.787879	6,056.494949
XGBoost	0.841606	42,916,496.732041	5,031.842606

Figure – 13 : Comparision Table

Requirements

Software Requirements:

- Jupyter Notebook
- Operating System : Windows
- Technology : Python 3
- Packages: Pandas, Numpy, Pyplot, XGBoost, Scikit-learn.

Hardware Requirements:

- Modern Operating System (Windows)
- 4 GB RAM
- 5 GB free disk space
- X86 64-bit CPU (Intel/AMD Architecture)

Future scope

Geometric Location Analysis:

Consider the impact of location on tourist growth by incorporating geometric location as a factor in the prediction model. Analyze transportation connectivity, city development, and proximity to other tourist destinations to gain insights into how location influences tourist demand.

Integration of Additional Social Networking Data:

Enhance the model's performance by incorporating more reliable social networking data sources. Explore data from platforms like Instagram or Facebook to capture user-generated content, sentiment analysis, and social engagement metrics, providing a broader range of insights for predicting tourist demand.

Prediction of Multiple Destinations:

We want to improve our model to predict the tourist arrivals for the multiple destinations.Due to lack of dataset avaliability we are restricting it to TTD only.

Collaboration with Tourism Stakeholders:

Engage tourism authorities, travel agencies, and other stakeholders to gather insights and incorporate their feedback. By collaborating with industry experts, ensure the model aligns with industry needs and can be effectively implemented in real-world scenarios.We may get the various type of data from the Tourism Companies.By using those data we can improve the accuracy of our model.

These potential future scopes aim to enhance the accuracy, performance, and practical applicability of your tourist demand prediction project, leading to better decision-making and improved visitor experiences within the tourism industry.

Conclusion

Our tourist demand prediction project for Thirumala Thirupathi Devasthanam (TTD) leveraged the Selenium framework for data scraping and implemented effective preprocessing techniques to create a high-quality dataset. Among the four machine learning algorithms tested, LightGBM emerged as the top performer, showcasing its superior accuracy in predicting tourist demand. This project highlights the importance of utilizing machine learning to optimize resource allocation, enhance crowd management, and improve visitor experiences at TTD. Moving forward, by refining the model and fostering collaborations with industry stakeholders, we can further advance tourist demand prediction for TTD, contributing to the growth and success of the destination.

References

[1] Rob Lawa, Gang Lib, Davis Ka Chio Fongc, Xin Han; Tourism demand forecasting: A deep learning approach, Elsevier-Annals of Tourism Research 75 (2019) 410-423.

[2] Anurag Kulshrestha, Venkataraghavan Krishnaswamy, Mayank Sharma; Bayesian BILSTM approach for tourism demand forecasting,Elsevier-Annals of Tourism Research 83 (2020) 102925.

[3] Tao Peng, Jian Chen, Chenjie Wang 2, And Yanshi Cao 1; A Forecast Model of Tourism Demand Driven by Social Network Data, IEEE Vol-9, 2021.

[4] Jian-Wu Bi a, Hui Li a, Zhi-Ping Fan b; Tourism demand forecasting with time series imaging: A deep learning model, Elsevier-Annals of Tourism Research 90 (2021) 103255.

[5] Karakitsiou, Athanasia, Machine learning methods in tourism demand Forecasting; MIBES Transactions, Vol 11, Issue 1, 2017.