National College of Ireland

Snowfall Prediction using Machine Learning Approach: Boston

MSc Research Project
Data Analytics

Atif Feroz Jeelani
Student ID: x17169992

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Atif Feroz Jeelani
Student ID:	x17169992
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	12/08/2019
Project Title:	Snowfall Prediction using Machine Learning Approach: Boston
Word Count:	8985
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Snowfall Prediction using Machine Learning Approach: Boston

Atif Feroz Jeelani
x17169992

Abstract

Snowfall plays an enormously essential role in the mountainous regions and lives of indigenous people which can affect the economic, socio-economic and livelihood of the people throughout the nation. It is a blessing but as well as a curse if a timely accurate prediction is not done. The problem lies in the accurate prediction of snowfall as the weather is a random phenomenon and its non-linearity is difficult to understand both in terms of complexity and technology. To this Machine learning is relatively robust to perturbations and is capable of understanding the non-linearity. In this research, machine learning algorithms (LSTM, Decision Tree, Random Forest and XG Boost) have been employed to yield better accuracy. Feature selection was done using Boruta algorithm. Experimental evaluations resulted in the highest accuracy shown by LSTM with an accuracy of 89.98% which outperformed other classifier models. Whereas, Decision tree and XG Boost resulted well in the overall performance of prediction with respect to other evaluation metrics. The results of this research added to the contribution of the knowledge in weather prediction in the domain of Snowfall for the machine learning industry.

Keywords Machine learning, Snowfall prediction, LSTM, Decision Tree, xGBoost, Random Forest.

1 Introduction

The seasonal Snowfall is an essential part of the Earths climate system. Snow cover is the major single constituent of the cryosphere which is covering an average of about 46 million square kilometers, yearly of Earths surface every year. Snow helps to regulate the temperature of the Earths surface and helps fill rivers and reservoirs in most regions of the world (Navid and Niloy; 2018). Snowfall has a significant impact on our society for better and for worse.(Srinivasan et al.; 2013) states that the accurate forecasting of snowfall has an important contribution to the growth of economy and safety consequences for the entire nation. A timely and accurate prediction of a snowfall event could mean the difference between a manageable snowfall and a city shut down by impassable road conditions. However, an accurate and timely prediction of weather has been one of the most dicult tasks across the globe (Thirumalai et al.; 2017). This is because of the fact, the weather is a random phenomenon which is non-linear in nature and its prediction has been very dicult both in terms of technology as well as complexity, basically (Mohapatra et al.; 2017). Snowfall prediction remains one of the difficult challenges in

weather forecasting for a meteorologist. Reason being that variables such as temperature, atmospheric pressure, humidity, wind which are main meteorological factors are dynamic in nature and generated in a non-linear fashion which affects the accuracy (Geetha and Nasira; 2014a). So a machine learning model is must needed which can understand the non-linearity of the data and increase in the accuracy of the snowfall prediction.

1.1 Motivation and Project Background

Snowfall plays an enormously important role in all the mountain regions due to the reason of snow accumulation and melt. This generates water resources due to snowmelt and glacial meltwater which plays a fundamental role in continental hydrologic cycle, industry and agriculture development of a nation (Xueting ZHANG; 2019). Most of the indigenous and artic people wholly depend on snowfall for their livelihood for thousands of years (Navid and Niloy; 2018).Snowfall gets accumulated on the ground which changes to snow cover giving rise to surface albedo, impacting the surface runoff. It is also closely related to local tourism. (Joshi; 2012) has acknowledged the fact that Himalayan snow cover is feeding source of major rivers in central Asia which supplies water to more than 1.2 billion people. On the other hand, extreme snowfall gives birth to a natural catastrophe like snow avalanches. Such disasters, toll human lives and damaged property worth millions every year (Tae Sohn et al.; 2009). An unexpected snowfall event adversely affects the livestock and ultimately leads to a decline in various economic and socio-economic sectors of a nation (Sejian; 2016). Though it plays a vital role in hydrology, glaciology, climate change, and agriculture but is also responsible in disaster management on the manifestation of avalanches and landslides due to heavy snowfall (Joshi; 2012).

From the past few decades, there has been an in-depth study and analysis of weather predictions. A wide range on the improvement has been made using machine learning techniques to predict a variety of meteorological applications like rainfall, earthquake, wind energy, solar energy, tornado, hail storm prediction and detection since late 1980s (McGovern et al.; 2017). However, in recent times there has not been much contributed in the domain of Snowfall prediction (Joshi; 2012).

(Joshi; 2012) has acknowledged the fact that very little amount of time has been spent in past few decades for developing a proper snowfall classification system. In recent years, few attempts have been made in statistical and numerical prediction in the domain of Snowfall. Problem lies in the accuracy of prediction (Mohapatra et al.; 2017). The present atmospheric state is sampled and the predicted state is computed using numerical solving equations. However, the traditional system of ordinary differential equations governing this physical model is not stable under perturbations and uncertainties in measurements of non-linear atmospheric conditions and incomplete understanding of this complex atmospheric process (Holmstrom and Liu; 2016). Also, current forecasts rely on observation from gauges, radars and satellites which occur at surprisingly narrow bands and smaller scales than which forecast zones can see and there exists extremely small temperature variance that defines boundary line between snow and rain (Navid and Niloy; 2018).

Modern analytical tools such as machine learning are relatively robust to perturbations and do not need to completely understand the physical process that governs the atmosphere. To tackle these problems a machine learning classification model is required to understand the non-linear trends. Deep learning models like LSTM (Long Short-Term Model) which is specifically designed for time-series has shown tremendous result

in weather forecasting (Chao et al.; 2018). But no work has been done in snowfall domain. For this research, the focus was to study and implement Supervised Machine learning classification models Random Forest (RF), Long Short-Term Model, XGBoost and Decision trees to tackle the problem. Later, comparison of classification model performance using evaluation metrics like Accuracy, Sensitivity, Specificity and Precision are taken into consideration to find the best performing model in the domain of snowfall prediction.

1.2 Research Question

***RQ**: Can we efficiently diminish and minimize the catastrophe that is caused to economic sectors and socioeconomic conditions by predicting/estimating the Snowfall considering the factors (temperature, humidity, wind, dew point, precipitation, sea level and visibility) using deep learning (LSTM) and other machine learning algorithms(Decision tree, Random forest and xGBoost) ?.*

The Research question addresses the problem of an accurate and timely prediction of Snowfall which affects the lives of people, especially living in the mountainous regions and lives of indigenous people in a catastrophic fashion which exploiting the economic and socioeconomic condition of a nation.

***Sub-RQ 1:** To what extent, can machine learning technique be used to understand the non-linear patterns of past weather data in order to increase the accuracy of snowfall prediction ? .*
To resolve this research question the following objectives are specied and employed.

1.3 Research Objectives and Contribution

The core objective to be investigated in this research work is to implement a deep learning model to tackle the research question and compare this model to the current existing successful models. This research work has been divided into stages of objectives as mentioned above.
Objective 1: A critical review and discussion of literature in the field of weather forecast and snowfall prediction.
Objective 2: To create a lag time series data followed by pre-processing and feature selection.
Objective 3: Exploratory Analysis of the snowfall data to find hidden relations between meteorological variables.
Objective 4 (a): Implementation, Evaluation and results of Long Short-Term Memory model.
Objective 4 (b): Implementation, Evaluation and results of Random Forest (RF) model.
Objective 4 (c): Implementation, Evaluation and results of Decision tree model.
Objective 4 (d): Implementation, Evaluation and results of eXtreme Gradient Boosting (xGBoost) model (Obj 4).
Objective 5: Comparison of the developed models.

Research contribution : This research is significant as it demonstrates the implication of a deep learning approach in the field of Snowfall prediction which has not been

implemented yet in this domain. In recent decades, few researchers have examined this critical and challenging weather prediction problem from the machine learning perspective. This study can replace the traditional method followed by mountainous regions to examine the non-linear patterns of meteorological variables and thus boost in accuracy of prediction. Timely and accurate Snowfall prediction can prove to be a boon for the welfare of people and the state economy.

The rest of the technical report is structured in the following fashion. Chapter 2 illustrates an investigation of the existing literature in the field of the weather forecast and snowfall prediction using different approaches. Chapter 3 demonstrates the scientific methodology followed for this research. Chapter 4 presents the implication and implementation of the developed models. Chapter 5 defines the results and evaluation of the predictive models and finally, Chapter 6 concludes the research study with a conclusion and potential future work.

2 Literature Review on Snowfall prediction and other weather phenomenon (2004-2018)

This literature review studies snowfall prediction in depth using various Statistical Techniques, Machine learning and Deep learning algorithm (Decision Tree, Regression Analysis, Random Forest, SVR, LSTM, ARIMA, ARMA, STL, K-Means Clustering, Baum-Welch algorithm, etc). The above section illustrates literature done by numerous Researchers in the above mentioned domain. In this section, thorough study of techniques has been done that are applied by researchers so that an idea is popped up about the model that can be implemented for this investigation. To showcase a better interpret ability of past work, we have divided this section into followings:

- Critical Review of Snowfall using Machine learning.
- Critical Review of Snowfall using Statistical Techniques.
- Critical review of other Weather Predictions using Machine Learning.
- Literature in reference to the Proposed Algorithms.

There is a good amount of literature already available in the weather prediction using various Machine Learning and Statistical Techniques. The following section gives a comprehensive overview in the relative domain.

2.1 Critical Review of Machine Learning Algorithms in Predicting Snowfall and Identified Gaps

There are mostly three methods that researchers usually incorporate to forecast snowfall, and these three methods include climatological snow ratio, neural network & lookup table. Whereas, Machine learning algorithms and predictive analysis always work in sync and therefore researchers have already experimented snowfall forecasting implementing various machine learning algorithms with the aim of preventing the detrimental repercussions of unexpected snowfall. This has been investigated by the researchers (Roebber

et al.; 2007), wherein they concentrated on forecasting real-time snowfall amount with the application of neural network. In order to for snowfall prediction researchers employed the application of above mentioned 3 methods in their analytical study that had surface temperature as its main feature. In order to carry out their research study, they gathered 53 snowfall reports in real time for the time period of 2 years (2004-2005 & 2005-2006) for the winter months ranging between Nov and March. Dataset selected for this study was taken for United States region in the east of Rocky Mountains and the chosen variables were provided as input to the applied neural network model which included features like precipitation, humidity, temperature and wind speed. The methodology incorporated by the researchers include snow ratio, neural network, and surface temperature based lookup table along with the application of standard verification methods (Bias, Mean, Median, RMSE). The results in this study, explained that in individual events the NN had performed better in terms of capturing the amount of snow. The reported results could have been improvised by incorporating certain improvements such that the predicted class probability could be explained within class. Also, there should have been some connectivity in the detail of predicted vertical motion and sounding profile which could have been refined and implemented on a large data which could have yielded a better accuracy score.

Whereas, identifying the factors that control the snowfall amount is another important area of research study that most of the researchers opted. With this aim in mind, the research study was conducted by (Xueting ZHANG; 2019). In order to predict the snowfall and the variables that influence snowfall, researchers implemented the use of MLR analysis and RF model. Researchers wanted to evaluate the important controlling factors of the snowfall in the region of Xinjiang of Tianshan mountains. The dataset was taken from 27 meteorological stations during winter months of October and April (1980 to 2015) and the variables chosen for the study include, temperature, relative humidity, longitude, speed, slope. The results illustrate that temperature and relative humidity influence the snowfall, with the significance of p less than 0.05, whereas low significance level was noted in the case of slope, wind speed and elevation. It was observed that random forest had a better performance in comparison to Multiple Linear regression which consequently indicates that non-linear trends would be more suitable in explaining the association between snowfall and predictor variables. In order to improvise the accuracy levels, researcher could have implemented standard verification method, namely, Mean, Root Mean Square Error, Median, etc. on the other side, if snowfall or rain rates are to be estimated on the higher altitudes, Deep Neural Networks works the best.

A Similar research was conducted by (Tang et al.; 2018). In his study he compared DNN results to the data gathered from Goddard Profiling Algorithm (GPROF). As the radars and ground gauges are with limited coverage at high altitudes, therefore to avoid this disadvantage, the data gathered for his study was used as training data, that provide direct precipitation observations from space that are CloudSat and Global Precipitation Measurement (GPM). Researcher used three different types of data sets for the research study that includes environmental data from European Centre for Medium-Range Weather Forecasts, passive microwave data from the GPM microwave imager, infrared data. These data sets are further trained to the reference data sets for the estimation of rain and snow. Once the analysis was run and results were generated, it was observed that DNN models outperformed the other model in both the testing as well as training periods. It was also noticed that infrared and environment data has the capacity to improve snowfall forecast incorporating DNN models. Finally, optimized DNN snow and

rain estimates were compared to ERA-Interim and Modern-Era Retrospective analysis for research, and its results depict that DNN based snowfall estimates outperform GPRO on a larger scale

2.2 Critical Review of Statistical Techniques in Snowfall Prediction and Identified Gaps

Many researchers opted for improvising the statistical techniques for the prediction of snowfall. Same was experimented in the following study wherein, researcher attempted to forecast heavy snowfall, and for this purpose he adopted a ternary forecast approach (Tae Sohn et al.; 2009). Author wanted to improvise the statistical modelling for forecasting. Researcher used daily observations on snow amounts and readings which he gathered from 17 different stations located in the Honam area in Korea over the time period of 5 years (2001-2005).For this research study, observations were considered using only winter months. In this research the target variable was a multivariate where there three ranges, less than 50mm, 50-150mm and above 150mm were factorized to 0, 1 and 2 respectively where the range 2 was considered to be a danger zone. Cluster analysis was implented for the analysis. On the basis of Model Output Statistics (MOS) which is a statistical method, two models, viz, three grade neural networks where a three level multiple logistic regression analysis was performed individually in order to attain high probability for prediction of snowfall. When the results were compared, interestingly the neural network performed no better than logistic regression model implemented and therefore concluded the preference of multi-grade logistic regression in the case of heavy snowfall for the region of Honam. Probably the neural network failed as the dataset was not large enough to be implemented otherwise it could have performed a better fit.

The rapid flow of snow over a mountainside or a hill is called an Avalanche which has lead to many causalities. In this regard it was (Joshi; 2012) who focused on the prediction of the next day values for the meteorological features by the implementation of statistical methods. In this research Multiple Linear regression (MLR) was employed for the hilly region of Ladakh and Manali (Patio). For the implantation of this method the dataset was used for only winter months for the duration of 10 years. In this prediction analysis 12 meteorological features were considered by incorporating MLR which were observed in the analysis for the previous day. Authors than ran a comparative analysis between observations computed through the regression model and forecasted parameters and recorded minimal variations between the two, and thus this model could prove to be a boon in the field of snowfall and avalanche prediction, consequently preventing its detrimental hazards. If researcher had taken more number of years and had considered the non-linear variations, this analysis could have yielded better results.

In order to predict quantitative snowfall, Hidden Markov Model was developed in the year 2017 by (Joshi; 2012) in two different ranges of Indian Himalayas, Great Himalayan Mountain Range and Pir-Panjal Range. Researchers have predicted snowfall for 2 locations in the Great Himalayan range and for the duration of only winter months ranging from the month Nov-April. In this research data has been recorded twice a day , at time 8.30 and 17.30hrs for the period of 40 years. This model was developed was implemented using the weather parameters such as min-max tempt, relative humidity, sunshine time, and wind direction for the duration of 20 years from 1992-2012. Two models were used for the prediction analysis of 2 days in advance and used Baum Welch approach for the optimization of Viterbi and feed forward algorithm model implemented in this research.

Researchers in this study validated the model for two winters (2012-2013, 2013-2014). The statistical tools that they used to compute the accuracy levels are as following, Percent Correct, Root Mean Square Error (RMSE), Critical Success Index, and Heidke Skill Score. The model thus formed forecasted snowfall in various parts of the Himalaya that was in sync with the observed one. Once the results were generated. In the end, results showed that the implemented model had increased in prediction value which is compared with the random prediction for 2 days in advance. RMSE showed improved results in terms of the optimized model when compared to the results for prediction of 2 days in advance. The researcher could however taken transitional probabilities in consideration while building the model for optimum results forming a HMM which is non-homogeneous.

Whereas for the prediction of seasonal snowfall, it is important to determine the capability of multiple discriminant analysis (MDA). Same was experimented in 2014 by the researchers named Daria Kluver and Daniel Leather. (Chatterjee et al.; 2018). The dataset that they considered was for 440 stations in US and it ranged from the time period of 1930 to 2006. This data was collected as a subset of US Historical Climatology Network, which was detailed as a high-quality data suitable for trend analysis by Kunkel et al.(2009). Researcher tested two variables in this study, viz, total snowfall and frequency of snowfall and the parameters which were incorporated for this model building were as follows, , Decadal Oscillation, Atmosphere Teleconnection Patterns, and EI Nino Oscillation, NAO, AO, Temperature, and Land Cover. Researchers had applied Jackknife analysis in this study and concluded that the Central United States, Ohio River valley, Great Lakes, and Upper Midwest regions depict the highest level of snowfall and they also highlighted the factors that impact decadal snowfall variation. Authors reported the results which depict that MDA forecasts was correct from 20%-80% in case of frequency of snowfall, whereas, in the case of total snowfall, it was correct from 20%-50%. In order to improve the results reported, researcher could have picked only those stations data that depict high forecast ability, consequently could determine the skill that could state if snowfall is due to model configuration or specific station attribute.

On reviewing the available research papers in the field of snowfall prediction, it could be noted that little amount of work has been performed for the prediction of snowfall. Therefore, to get a better understanding on the applied machine learning perspective of weather prediction like rainfall, in order to implement a better approach yielding high accuracy for in the field of snowfall.

2.3 Critical Review of Machine Learning Techniques in Predicting the Other Weather Phenomenon

On the other hand, agricultural economy is completely dependent on rainfall and crop productivity, therefore this makes farmers depend on the rainfall for productivity of the crops. This is why it becomes so essential to accurately predict the rainfall so that the crop produces and resources are effectively utilized. Regression techniques are known to yield a decent level of accuracy in this regard was analyzed and confirmed by (Navid and Niloy; 2018). They studied the use of multi-linear regression in predicting of rainfall. The main focus of the study was to evaluate the prediction of rainfall with the application of multi-linear regression. The researchers gathered the data over the band of 30 years from Rajshahi, Bangladesh and for this purpose they implemented precipitation, cloud cover, vapour pressure and average temperature as the predictors and further applied multiple regressions on the collected data set. Researcher reported predictable equation

between rain and various other factors considered. Consequently when MLR equation was implemented with the test data, it was observed that there was actually a very minor difference between actual and predicted rain amount, thus confirming the model, despite omitting the other factors that could influence the rainfall.

Whereas (Kala and Vaidyanathan; 2018) attempted the model based on Artificial Neural Networks for prediction of rainfall. The researchers implemented Confusion Matrix and Root Mean Square Error on which the prediction accuracy was based. They yielded the acceptable accuracy level through the analysis, thus validating the model. Although their model generated acceptable accuracy levels but could have been improved by running a comparison various classification algorithms and also collecting larger amount of data for the analysis.

Also, in the domain of air temperature, weather dataset has been trained with deep learning algorithms. It was in 2015 when a model was formulated for the same in the North Western region of Nevada by (Hossain et al.; 2015). Researchers in their study compared the performances of Stacked Denoising Auto-Encoders (SDAE) with Artificial Neural Networks (ANNs) in order to forecast air temperature and therefore confirm as to which model performs better. They computed raw sensor data for the time period of 1 year and considered only 4 variables for the analysis that includes barometric pressure, temperature, wind speed and humidity for every hour taken from only one weather station. Results reported 97.97% accuracy where authors used cross validation on test sets. SDAE along with good choice of variables, whereas, results generated 94.92% accuracy level where traditional ANN was implemented. Thus, it was concluded that Deep neural network with SDAE outperformed Multi-layer feed forward network (ANN). Whereas, other rainfall prediction studies confirmed radial basis function neural network to be the most suitable technique for the same as it illustrated the least training error, whereas, other techniques, such as back propagation of neural network and the model with GRNN was over fitting the data. Thus rejecting the two (Tharun et al.; 2018). Also through the study carried out by (Dubey; 2015) it was reported that, results generated through RBFNN were more accurate than K-means Clustering and Artificial Neural Networks.

2.4 Critical Review of Literature in Reference to the Proposed Algorithm

Snow can hamper flight landings, take-offs, or even a smooth flight journey as well. Huge chunks of snowfall make it really difficult to keep runways clear. Thunder snow storms and blizzards can cause icing, visibility or turbulence issues during landings and flight travels. Therefore predicting snowfall earlier is of utmost importance even in the domain of flight navigation as well. This was taken up by (Aftab et al.; 2018) in their research study wherein they wanted to have a comparison of visibility forecasting using LSTM and ARIMA models. They selected Hang Nadim Airport of Indonesia for their model for predicting the parameter of visibility as the target variable which was combined with different weather features like humidity,dew-point and temperature. Once the implementation of models was done they compared the RMSE values generated from both the models and concluded that Long Short-term Memory model yielded high accuracy compared to ARIMA in time series analysis in both the cases of values generated.

In, another technical research which was carried by (Geetha and Nasira; 2014b), the author depicted the performance of a decision tree based model for the prediction of weather activities like rainfall, thunderstorm, fog and cyclone. Interestingly, this model

resulted with the accuracy of 100% which should not have been considered as the data was highly biased and the author could have relied on sensitivity. Also, in this research random forest proved to be better than Support Vector machine and Decision tree models.

Similar investigation was conducted where a research presented the performance of deep neural network architecture which was implemented on a time-series data for prediction of weather, for this purpose(Zaytar and El; 2016) analysed the performance of a 3 layered LSTM model. For this research the author used the hourly based data which ranged for the duration of 15 years, And the aim was to implement deep learning model in the city of Morocco for predicting weather parameter for next 24 and 72 hours. The results obtained made it clear that LSTM yielded far better results to the compared traditional neural network approach and proved to be a better alternative to forecast weather conditions.

In order to prevent the disasters created by the floods, an accurate and timely prediction of rainfall is required which can result in minimum possible stochastic error. With this aim in mind,(Chao et al.; 2018) came up with a model (STL) which decomposed the referral time series into trend, season and the residue applications for sensors of MEMS which was able to tackle the problem in real time rainfall prediction in Wuhan. Once the trends were received from the observed series, they were compared with the trends that were generated from the observed authentic dataset. The results suggested that MEMS sensors are believable. Researchers in this study have also used LSTM for predicting real time rainfall which is based on the data observed, it is then compared to a Random Forest, BPNNs, Support Vector Machine, Moving and Auto-regressive Average. Once the results were reported, it was observed that deep learning model (LSTM) showed better results when compared to SVM, RF, and BPNNs in Seasonal time as well real time predictions. It was concluded that LSTM is powerful enough in extracting the changed rules of seasonal rainfall by which we could be optimistic of its performance in the field of snowfall domain as well which can replace the old and expensive traditional approach and can eventually be the optimal benchmark that could be used in snowfall prediction .

2.5 Conclusion

On the basis of literature reviewed and identified gaps, it is quite evident that there is a need to develop a model which can predict Snowfall and answer research question. The models used by the researchers in the above literature include, regression techniques, deep learning algorithms, artificial neural networks, statistical techniques, etc. This review gives a clear vision of understanding related to which models can be employed in order to attain high accuracy in the domain of Snowfall. The models that are implemented are Random Forest, Long Short-Term Memory (LSTM), Decision Tree and xGBoost as a classifier. Based on the literature review, Random Forest technique as a classifier was proved to be better in terms of accuracy for rainfall prediction, which will be the base model of this technical report. From the review it can also be concluded among the time-series analysis, Long Short-Term Memory model performed better comparatively. Also, there was a paper that suggested Decision tree to be effective enough for predicting rainfall with the resulted accuracy of 100%. Therefore all this has persuaded me to implement these models in my analysis so that better results could be yielded.

3 Scientic Methodology Approach Used, Data Preparation and Project Design

3.1 Introduction

This section illustrates the techniques and the scientific methods followed to execute this project. This includes the transliteration process ow carried for implementation and also includes a technical design of implementation which consists of a two-tier structure. Modified Knowledge Discovery Databases (KDD) data mining method has been used as it fits best in my scenario. All the stages of KDD have been explained in this section along with visual analytics which provides data insights.

3.2 Snowfall Methodology Approach

In the field of data analytics, all the researches are carried out by any of the three methodologies viz SEMMA, KDD or CRISP-DM. In this research, modified KDD (Knowledge Discovery Data) is adopted in order to perceive optimal knowledge from dataset. This method suits the project scenario as the deployment of models in the business layer is not applicable in KDD. It was first introduced by Fayyad et al. (1996) who explained it as non-trivial, novel and potentially useful for identifying ultimate patterns in data. It comprises of 5 stages which when adopted in this research can be seen in Figure 1.

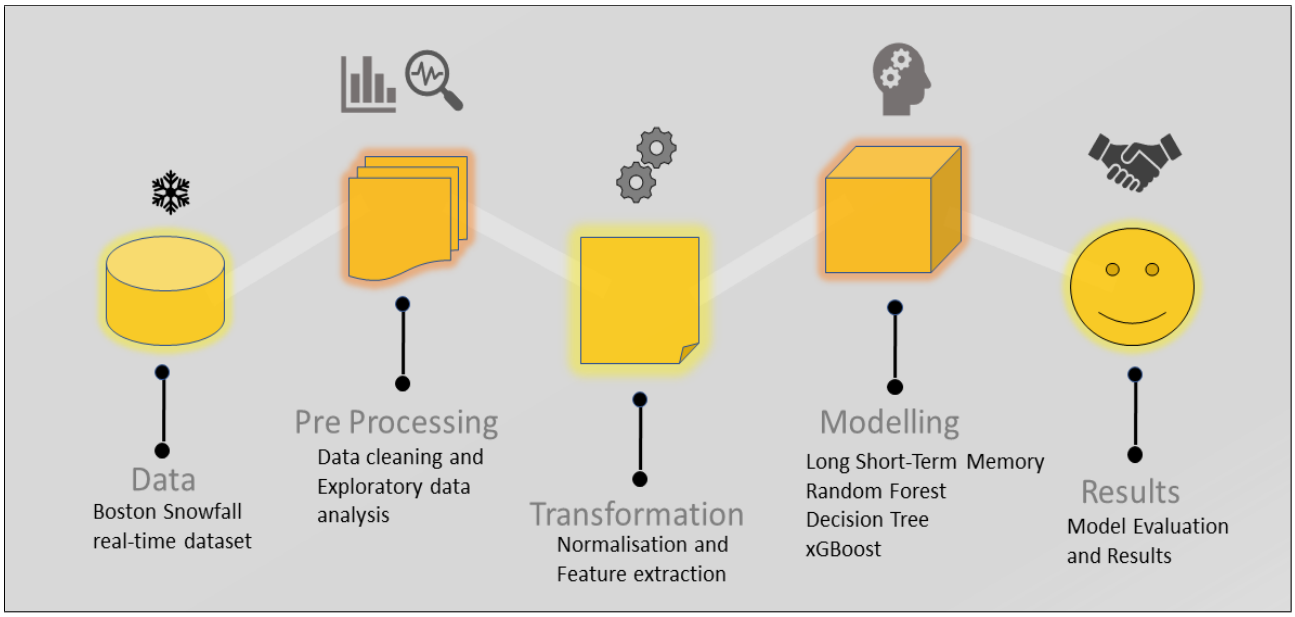

Figure 1: Snowfall Methodology Approach

3.2.1 Data selection and Understanding

In order to achieve the accuracy of a snowfall prediction model, it is very important to capture the right meteorological factors. According to (Zhang et al.; 2019), relative humidity, temperature, and latitude were identified as the most important factors in predicting snowfall. Therefore, for conducting this research, data was accumulated for

the region of Boston, Massachusetts with the dataset comprises mainly of 6 meteorological variables humidity, temperature, sea level pressure, visibility, wind, and precipitation. It also gives us readings for the total amount of snowfall and event occurred (Snow/Rain). The dataset comprises of 10 years of historical data (2008-2018) for which daily data is taken. The data comprises of High, Low and Average of each variable. All the data is publicly available in its entirety on Weather Underground. This collation of data was taken from the Kaggle website over a period of 10 years from 1 weather station of Boston. It contains 21 meteorological features for this duration recorded on a daily average value. It is of worth stating that the data used in this study is currently most updated in the snowfall domain

3.2.2 Data Pre-processing

Data pre-processing is an essential step before implementing any model. It is compulsory that trivial information of the data needs to be eliminated in order to obtain better results from the models. The Boston snowfall dataset was extracted from the website was in the raw form and was checked for missing values and NA values. The data consisted of little outliers and special characters like percentage, brackets, comma which was removed in R using G-sub function. After this, the names of the columns were renamed for the better understanding of the data and the unwanted attribute was removed followed by column selection for analysis and aggregation of data. This was followed by Exploratory data analysis which is important to gain insights of data and progress to the next stage of the investigation

Data Exploratory Analysis :
In this stage the snowfall dataset is explored to gain hidden knowledge from the real-time dataset. It is an important step before modelling to gain value insights of data which is considered while modelling. In this research several exploratory techniques were implemented using R and SPSS. First we ran a MLR analysis to analyze which variable shows maximum variability to the dependent variable. It can be seen in figure 2:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.549[a]	.301	.300	.253	.301	268.436	6	3742	.000

a. Predictors: (Constant), Wind, Visibility, Temp, Sealevel, Humidity, DewPoint

Figure 2: Model summary

In the model summary of the regression analysis we get R value as 54% which explains the quality of prediction of the dependent variable. Here we notice R square value to be 30% which explains the variance in the dependent variable that is explained by the features.

Coefficient table can be seen in Figure 3. Coefficient Beta value explains how uniquely the features has explained the dependent variable. Here we can see temperature has explained the highest variance in the dependent variable followed by Dew point with Sig. value less than 0.05 which makes it statistically significant and least variance is explained by wind. This analysis depicts the importance of Temperature and Dew point to be included in the model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	5.038	.601		8.387	.000	3.860	6.216
	Temp	-.014	.002	-.811	-7.729	.000	-.018	-.011
	DewPoint	.008	.002	.467	3.868	.000	.004	.011
	Humidity	-.004	.001	-.188	-3.970	.000	-.006	-.002
	Sealevel	-.129	.020	-.102	-6.562	.000	-.167	-.090
	Visibility	-.046	.003	-.303	-16.040	.000	-.051	-.040
	Wind	.002	.001	.027	1.761	.078	.000	.005

a. Dependent Variable: Target

Figure 3: Coefficients Multiple linear regression

After this, correlation plot was drawn to check the correlation between variables. In the correlation plot, the blue color represents the variables with positive correlation and the red color represents the variable which are inversely correlated. The dark shade illustrates the strength of the correlation between different features is seen in Fig 4.

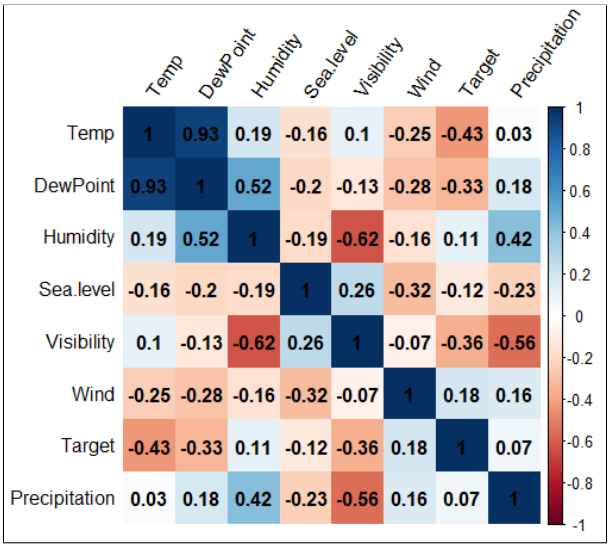

Figure 4: Correlation Plot

From the illustration, it is quite evident that humidity shows the max correlation to target variable followed by wind and precipitation while as temperature and dew point is highly inversely correlated to the snowfall.

Feature selection can be seen in figure 5, which was done before modelling which is an essential part while exploring the data. For this Boruta algorithm was implemented in the initial stage which depicts the predicting capacity of the predictors that show variance to the target variable. It can be seen the algorithm shows factors like Temperature , Humidity and Dew point explaining highest variance to the target variable. The information gathered from this data exploratory analysis gave us a clear picture about the variables and the models were built accordingly.

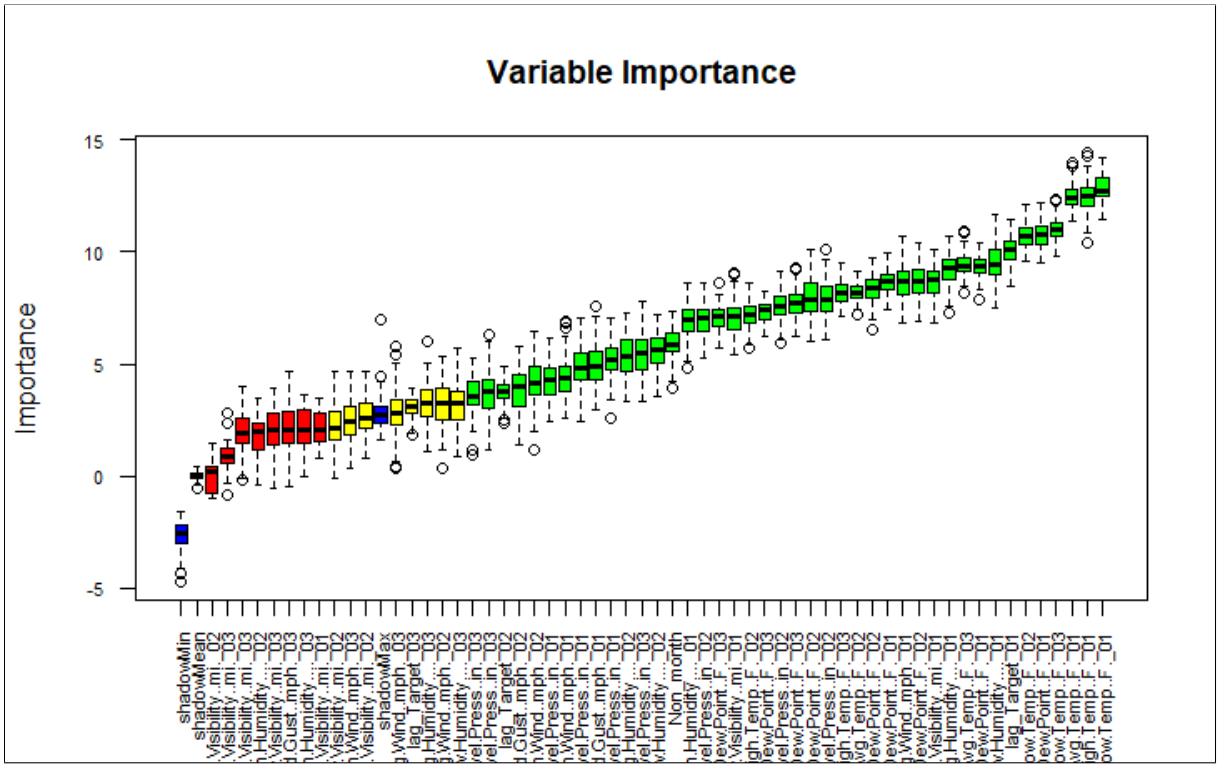

Figure 5: Variable importance plot

3.2.3 Data Transformation

In this stage of data pre-processing, all the obligatory transformations were carried out for the snowfall dataset in R. Data is transformed using methods like factorization, dimension reduction, normalization, feature extraction, to make it more suitable dataset which is essential before implementing machine learning model. For this dataset, the dependent variable was the Target variable which had 3 levels snow, rain, and nothing. This was transformed into a binary factor of snow or nothing and encoded as 0 or 1. The variables like year month and date were changed into factor type from int. The data were normalized to avoid biasing. In the end, feature selection was carried to avoid the least significant predictors which did not show much variance in the dependent variable and was removed from the analysis.

3.2.4 Data Mining

This stage comes after complete data pre-processing. In this stage, the data is ready to be modeled. Here the dataset was divided into test dataset and train dataset so that test

data can be used to test the performance achieved by the model on the trained data. In this investigation, we trained the algorithms using train data and evaluated it on the test data for which loss function was calculated to know the difference between the actual values and predicted values. This explains how well our train data effectively works on unseen data. Taking in consideration the past snowfall and rainfall implemented reviews, a deep learning technique Long short-term memory (LSTM) was implemented on the time series snowfall data with other machine learning algorithms named Random Forest (RF), Decision tree and Extreme Gradient Boosting (XGBoost).

3.2.5 Metrics and Criteria used for Evaluation

It is essential to estimate the performance of the implemented model for snowfall prediction to know about the reliability of the model. We use the unused test data to give us an estimated future performance. The forecasted values are compared with the actual values of the test data by calculating the error value which determines the quality of the model. The performance of the used models was evaluated using four evaluation metrics, precision, accuracy, recall, specificity, and F1 score.

Accuracy - This evaluation metric is simply a ratio of the correctly predicted observation of the total number of observations. Accuracy is a good measure when the dataset is symmetric. Therefore, in our case, we check with other performance models as well. (Alnoukari and El Sheikh; 2012):

$$\text{Accuracy} = \frac{tp+tn}{tp+fp+fn+tn}$$

Figure 6: Accuracy formula

Precision - It is the ratio of calculated true positives to the total correctly predicted false positives and true positives. This metric is also named as predicted value (PPV) (Alnoukari and El Sheikh; 2012). Mathematically :

$$\text{Precision} = \frac{true\ positives}{true\ positives+false\ positives}$$

Figure 7: Precision formula

Recall - It is also termed as Sensitivity. This True positive rate is calculated by the ratio of correctly predicted positive observations to the ratio of all in the actual class. Simply, as number of actual positives which were mis-classified. (Alnoukari and El Sheikh; 2012). Mathematically :

$$\text{Recall} = \frac{true\ positives}{true\ positives+false\ negative}$$

Figure 8: Recall formula

F1 Score - It is calculated as the weighted average of both Precision and Recall. Therefore, takes both false negative and false positive in account.

Specificity - This metric of evaluation can be defined as the portion of actual negatives which can be calculated by (True Negative)/(True Negative + False Positive).

3.3 Design Specications and Architectural Design

In order to develop a robust and secure model for predicting snowfall, an architectural representation of the developed model is given in figure 9. The diagram represents the technology, tools and techniques used in this process and illustrates the order in which each stage has been implemented to generate better accuracy for the snowfall prediction. The design also has connectivity with modified KDD methodology

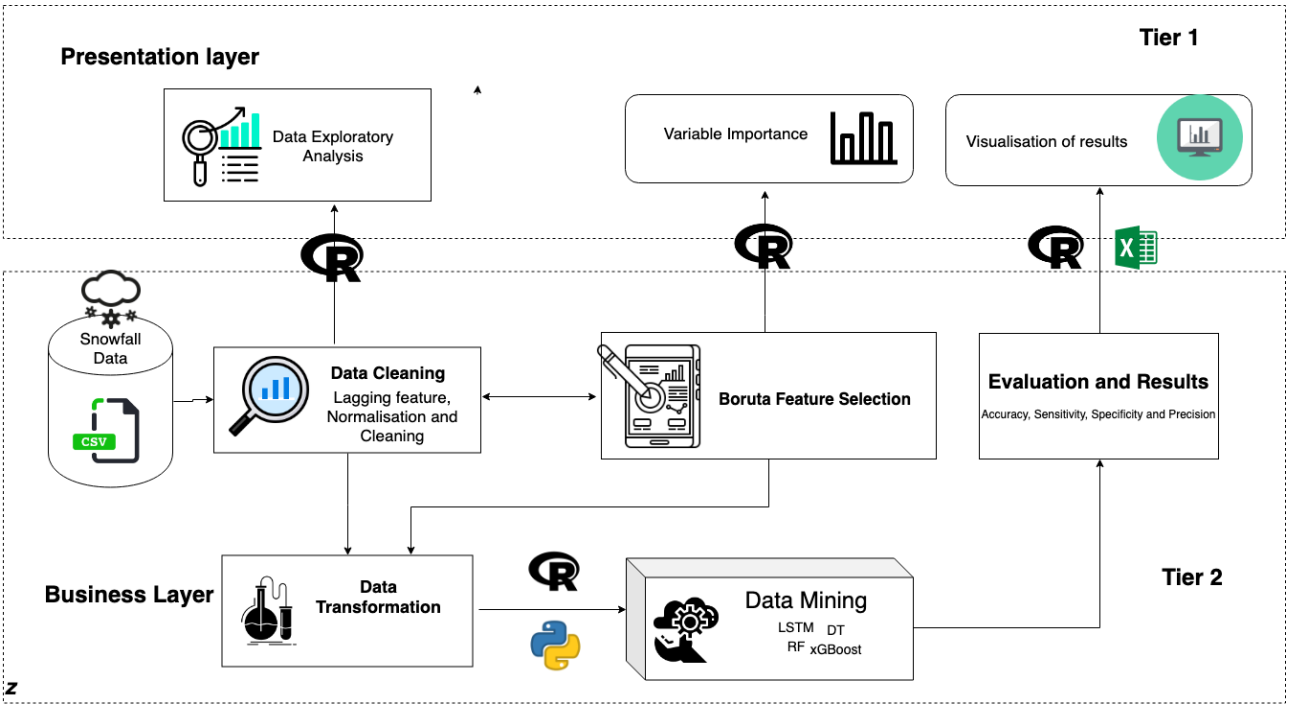

Figure 9: Architectural design of Snowfall prediction

The architectural design follows a two-tier design which includes two layers, presentation layer (Client tier) and application layer (Business Logic tier). In the business logic tier the selection of data in raw form takes place which is then followed by data preprocessing in the R tool where data is cleaned, and exploratory analysis of data is carried for better understanding. This is followed by feature selection and transformation of data. In this layer training of models and evaluation of results are carried. The tier 1 or the Presentation layer only comprises of the visual analytics for the client side undersatnding.

3.4 Conclusion

Hence, we follow Snowfall methodology approach for this project as per it fits the best. The process flow diagram explains how data analyst can approach this problem. The project flow diagram is used in chapter 4 for the implementation of prediction models. Further, the useful insights from data exploratory analysis by plotting correlation matrix and multiple linear regression were used in order to build the models.

4 Implementation, Evaluation and Results of Prediction Models for Snowfall Prediction in Boston

4.1 Introduction

This chapter of the technical report provides the entire process of employing and analysing the classification models developed for Snowfall prediction. It also contains the technical specifications that were considered for the execution and regarding the implementation of the machine learning algorithms (LSTM, Decision tree, Random Forest and xGBoost). This section delivers the information about the technical configuration including tools used and packages. The motive is to check the best fit model to obtain high accuracy.

4.1.1 Technical Configuration

Table 1 provides the technical configurations of the system, the model packages used and the tool used for the deployment of the model.

Table 1: Configuration and Specification table

Approach	**Decision Tree**	**Random Forest**	**XG Boost**	**LSTM**
Programming tool	R	R	R	Python
Machine RAM	8GB	8GB	8GB	8GB
Model packages	rpart	randomForest	xgboost	lstm
Processor	Intel i5 2.20GHz	Intel i5 2.20GHz	Intel i5 2.20GHz	Intel i5 2.20GHz

4.1.2 Implementation

The Snowfall dataset was obtained for the region of Boston, Massachusetts. It was extracted in a raw .csv format. This real-time data is obtained from General Edward Lawrence weather station with over the period of 10-years with (Latitude 42.3642998, Longitude -71.0052033). The time period ranges from (January 1, 2008) to (August 4, 2018) and contains daily data. The dataset comprises of 7 different meteorological parameters (temperature, humidity, wind, dew point, precipitation, sea level, and visibility) which are taken into consideration to predict snowfall[1]. The experiment was carried using programming tools R language and Python, as well as MS Excel. R studio provides better packages for our base models whereas Keras is used as a platform to implement deep learning model. The first and important stage of pre-processing was creating the lag time series in our dataset[2]. Considering the fact dataset is being used to predict snowfall, the lag shift has been taken for 3 days (n- 3n) to increase the probability of accurate classification. After this, all the missing values were checked. Target variable was multivariate (rain, snow, and nothing) which was transformed to bivariate (snow = 1, nothing = 1) followed by normalization of all the features using Caret package. Special characters were removed from the data using G-sub function and the column names were renamed in Excel for better understanding. Boxplot was generated using boxplot() function to check the outliers which depicted very few outliers which were not removed. In

[1] `https://nsidc.org/cryosphere/arctic-meteorology/factors_affecting_climate_weather.html`

[2] `https://https://www.business-science.io/timeseries-analysis.html`

this case, very high or low temperature or humidity plays an important role in explaining the variance of the dependent variable. After all the pre-processing, feature selection was done using inherently binary classification algorithm, logistic regression which tries to find the best hyperplane in K-dimensional space[3]. After feature selection was done 20 top features were selected which were uncorrelated and non-redundant to improve the model. Before training the classifiers, the dataset was split into training and test data using caret package in R. The data was split into 80% training and 20% for testing. For each model, K-fold cross-validation was performed in order to avoid overfitting of data where K was set to 10 folds for every model. In the end, the confusion metrics were drawn for every model and their AUC curve was plotted which is further mentioned in detail in the section (4.4). In the end, the prediction of the trained data was tested against the test data (unseen data) using the caret package. Different metrics were are used to know how significantly the model is performing.

4.2 Implementation, Evaluation and Results of Decision Tree Model

Decision tree is a class of very robust and powerful Machine Learning model which is capable of attaining high accuracy in both classification and regression problems while being highly interpretable. The knowledge achieved by decision tree by training is directly formulated in a hierarchical structure which is a graphical representation of solutions based on certain conditions. It contains the root node, internal node, and leaf node.

Implementation Process: Decision tree was the first model to be implemented. Among all the meteorological features top 20 features from the variable importance were selected which was used by decision tree for classification. To start with the training process of the decision tree, the dataset was split into training and test where 80% is used for training the model and 20% data is used for the test set. Lag for every engineered feature for 3 days has been taken into consideration to increase the probability of an accurate classification using mutate() function. Decision tree model was applied using rpart() function followed by k-fold cross validation where k=10 (10 fold) to avoid overfitting of data and accurate results. At the end summary of decision tree was obtained and decision tree plots were made using rattle package for fancyRpartplot() and rpart.plot package for prp() function which can be seen in Figure 10.

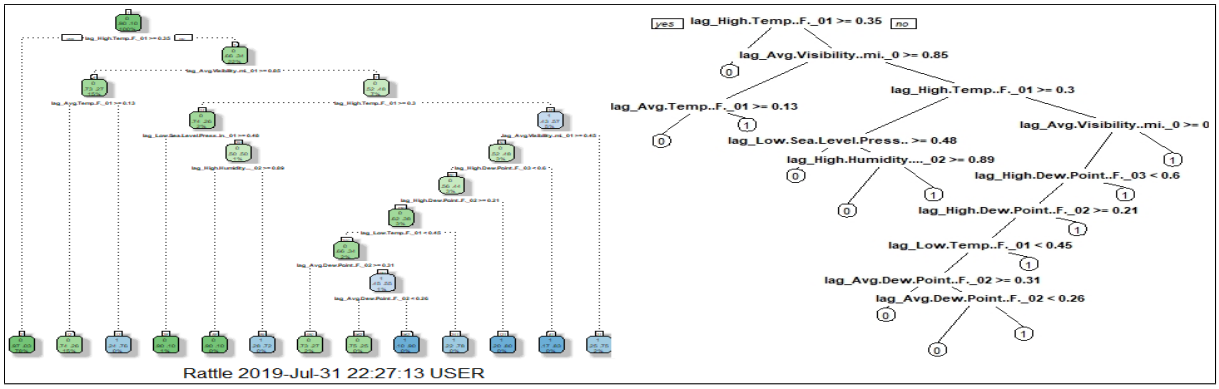

Figure 10: Decision tree for snowfall prediction

[3] https://https://towardsdatascience.com/model-based-feature-importance-d4f6fb2ad4031

4.2.1 Evaluation and Results of Decision Tree Model

The Decision tree model results can be seen in table 2. It was evaluated on the test dataset for 3 years. For the evaluation of this model, a confusion matrix was obtained. Evaluation metrics like Accuracy, Sensitivity, Precision, F1score, and Specificity were addressed. From the table, it can be clearly seen that the Decision tree model correctly predicted snowfall with 80.77% accuracy measure which explains the overall snowfall prediction. The model explains the Sensitivity and Specificity with a good score of 78.38% and 81.03% respectively. Precision and F1 score were observed to be 31.18% and 44.61%. The aim was to achieve a high sensitivity (recall) measure which gives us the performance of the classifier which has been achieved. These were the results obtained for all the engineered features.

Table 2: Results of Decision Tree

Model	Accuracy	Sensitivity	Precision	F1 Score	Specificity
Decision Tree	80.77	78.38	31.18	44.16	81.03

Also, a Receiver Operating Characteristic (ROC) curve was plotted which shows the true positive rate and explains how much the model is capable of distinguishing between classes which can be seen in figure 11.

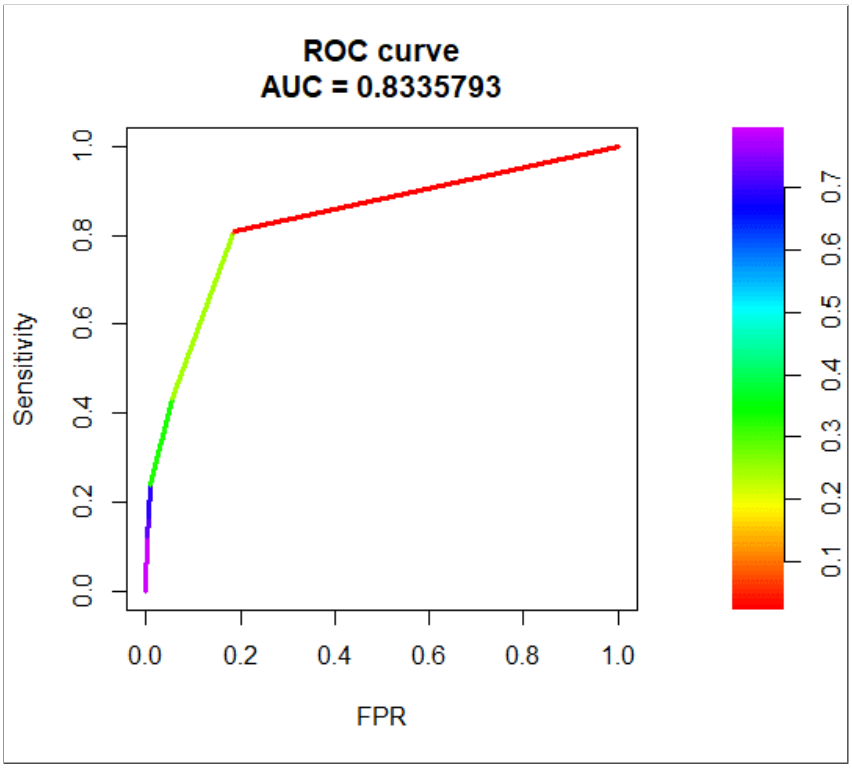

Figure 11: ROC curve for Decision tree

The more the curve towards top-left corner the better the result which can be seen from the figure. From which we get AUC value of 83.35%.

4.3 Implementation, Evaluation and Results of XG Boost Model

eXtreme Gradient Boost algorithm is the employment of gradient boosted decision trees which increase predictors successively and rectifies the previous models. This model combines the numerous trees generated by tree-based models which are having low accuracy in order to create a model with high accuracy. After every iteration weights are not assigned to the classifiers by this model, instead it matches the new residuals to the new model of the previous prediction and then enhances the latest prediction while reducing the losses.[4]

Implementation Process: For the implementation of XG Boost model, all the selected top 20 climatological features were considered which were feature engineered. All the selected important climate features for which lag of 3 days was taken were used in the development of this model. The data was split into 80% as training and 20% as test data. xgboost package was used in R studio to implement this model. For tuning the parameters Grid search technique was used. Control parameters were setup for optimization of the model, values such as gamma, nrounds depth of the tree, subsample were set. In order to avoid overfitting of data for optimal results, the model was cross-validated using 10-fold cross-validation. At the end, a summary of xGBoost was plotted using DiagrammeR package. xGBoost also gives the feature importance plot as an output from where predictorswith high variability can be observed in Figure 12.

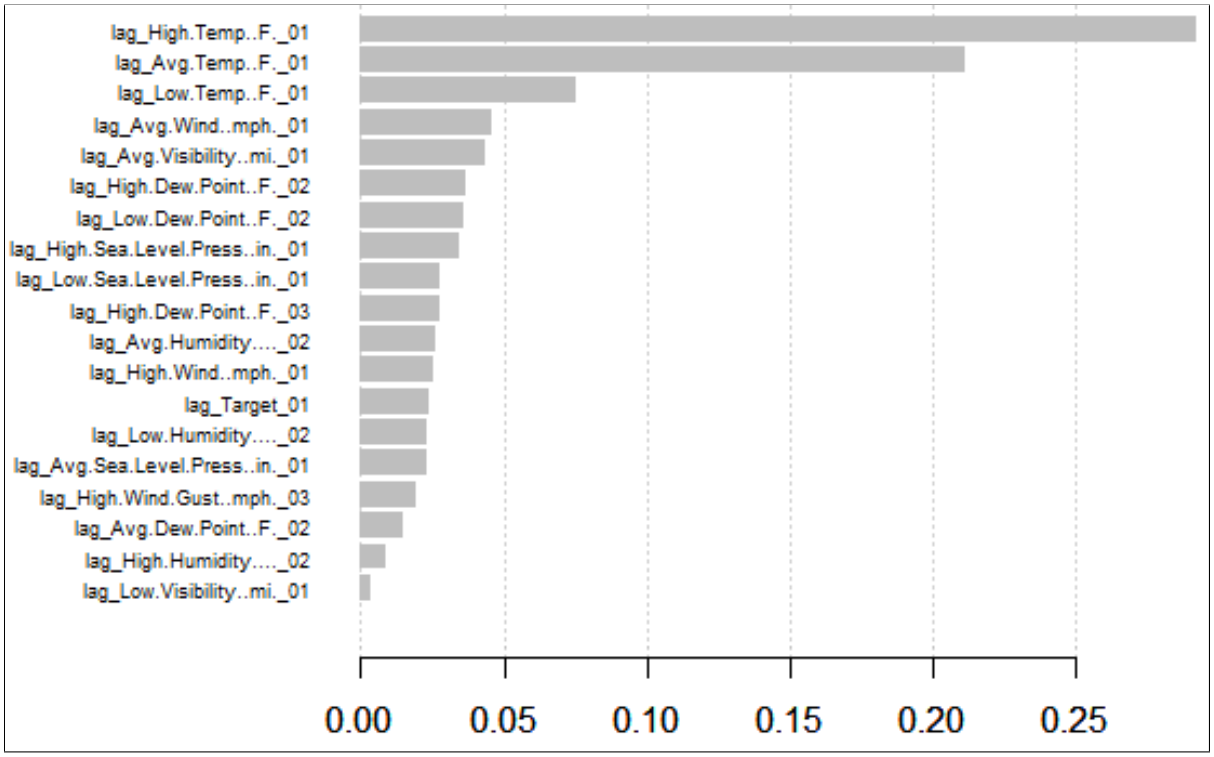

Figure 12: xGBoost variable importance graph

[4]Like this one: `https://machinelearningmastery.com/gentle-introduction-xgboost`

4.3.1 Evaluation and Results of eXtreme Gradient Boosting

XG Boost results can be seen in table 3. This model was evaluated on test data for the duration (2016 - 2018). After the deployment of the model, the confusion matrix was obtained. Evaluation metrics like Accuracy, Sensitivity, Precision, F1score, and Specificity were addressed. xGBoost correctly predicted overall snowfall event with an overall accuracy of 81.04% which can be seen in the table. The developed model explains the metric of Sensitivity and specificity with a score of 79.97% and 81.93% respectively. High sensitivity explains how well the model performed in predicting snowfall events when it actually snowed which makes the model significant. Precision and F1 score were observed to be 30.68% and 43.20%. Overall, this classifier performed well in predicting snowfall.

Table 3: Results eXtreme Gradient Boosting

Model	Accuracy	Sensitivity	Precision	F1 Score	Specificity
XG Boost	80.04	72.97	30.68	43.20	81.93

At the end, a Receiver Operator Characteristic (ROC) curve was also plotted as it reflects on the diagnostic ability in the case of binary classification. From figure 13, it can be illustrated that XG Boost has shown a tremendous measure of predictive accuracy. XG Boost observes 96.88% area under the curve which explains how much model is distinguishing between the classes.

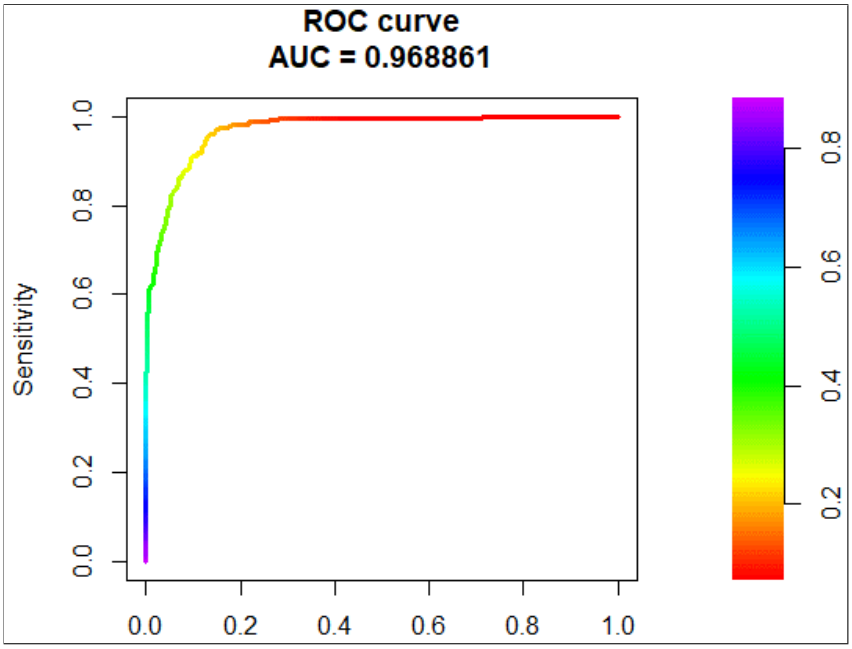

Figure 13: xGBoost variable importance graph

4.4 Implementation, Evaluation and Results of Random Forest Model

Random forest algorithm is a mixture of trees that in this case is used as a classifier for predicting snowfall. The random forest as a classifier is a mixture of multiple classifiers that is a tree structure working on the principle of decision trees. It creates a number of specified branches which is specified while implementing the model to avoid overfitting of data. This algorithm is designed in a way that makes it very flexible and easy to use [5]. It gives promising result even without any hyper tuning parameter. (Xueting ZHANG; 2019)

Implementation Process: The next tree-based model was Random forest which was implemented using R studio. For the execution of this model same procedure was followed the data was split in 80 20 ratio for training and test respectively using lag values for up to 3 days. All the important features were only selected to train this tree-based model to get an optimal result. Random forest was implemented in R using randomForest package. This model was optimized with the help of tune length available in Caret package. 10-fold cross validation was used in this model as well. Random forest also gives feature importance plot which was plotted using varImpPlot() function which can be seen in figure 14.

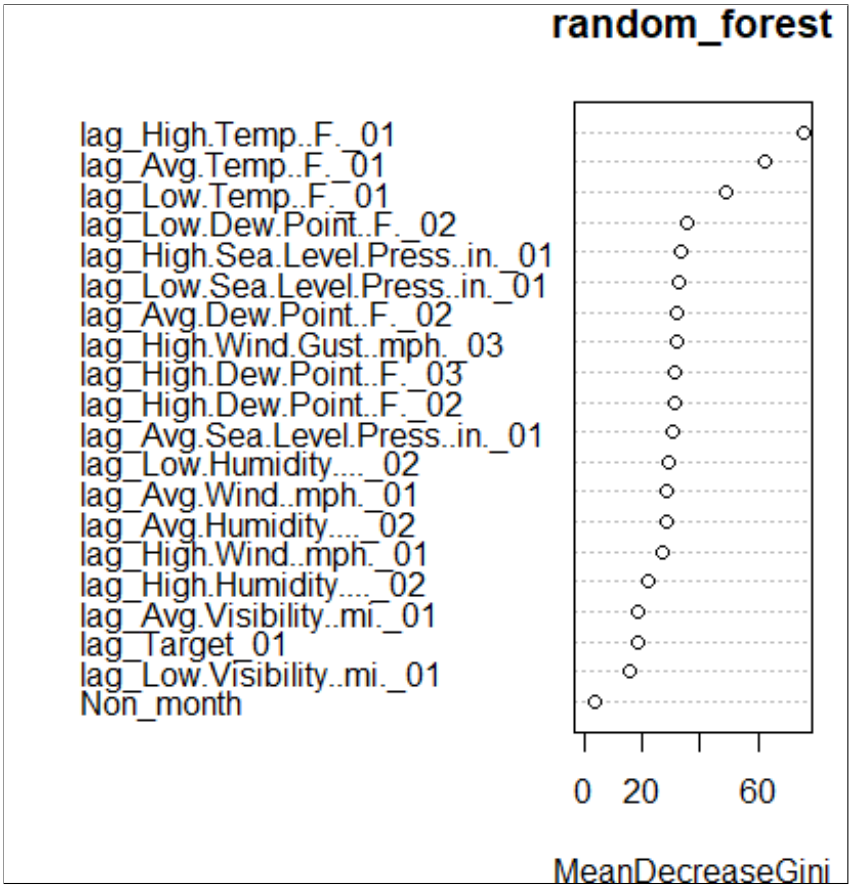

Figure 14: Random Forest variable importance graph

[5] Like this one: `https://towardsdatascience.com/the-random-forest-algorithm-d457d499cd`

4.4.1 Evaluation and Results of Random Forest

This tree-based model was also evaluated on test set for 20% of the data. Cross-validation was used for evaluation to avoid overfitting and observe the optimal scores for every evaluation metric. Random forest did not show good performance in predicting snowfall event as compared to other developed models. The accuracy percentage is 78.24 which is comparatively low. But however, Random forest has successfully predicted the snowfall events when it actually snowed which is interestingly the highest sensitivity score attained by the developed models with 81.08% making the model significant in predicting snowfall. On the other hand, good specificity of 77.92% is achieved which explains how well the model was in predicting when it was not snowing. Also, precision and F1score is taken into consideration which recorded 28.70 and 42.4 percent seen in table 4.

Table 4: Results of Random Forest

Model	Accuracy	Sensitivity	Precision	F1 Score	Specificity
Random Forest	78.24	81.08	28.70	41.4	77.92

4.5 Implementation, Evaluation and Results of Long Short-Term Memory Model

Considering the fact the obtained Snowfall data is a time-series data, LSTM has been implemented for the classification of snowfall as it specifically designed for time-series and can handle sequential data. The meteorological variables like humidity and temperature are non-linear in nature and the proposed non-parametric algorithm can very well capture this non-linearity. LSTM is an upgraded form of recurrent neural network (RNN) overcoming the shortage of gradient problem which occur in training the deep neural network (Zaytar and El; 2016).

Implementation process: To implement a deep learning model, LSTM was used as a classifier. For the development of this model Python 2.7 was used as it provides a variety of packages and libraries for LSTM. Keras was used as a platform to implement LSTM. First of all, we imported libraries like NumPy, Matplotlib, and Pandas for calculations, graph plotting and data manipulation respectively. Then the lag was taken up to 3 days for every feature. For feature scaling, MinMaxScaler was used from sklearn metric packages to achieve optimal performance. Data was split to (80 20) ratio for test and train set. Stacked LSTM was implemented for this case of classification where three-dimensional input layer is used for prediction using Dense function. Activation functions were kept to default where sigmoidal was used as activation for inner cells and dropout value set to 0.3 which means 30 percent of the layers were dropped. At the end confusion matrix and the diagnostic plot is plotted.

4.5.1 Evaluation and Results of Long Short-Term Memory

This deep learning model was also evaluated using 20 percent test data. 10 fold cross-validation was also implemented in LSTM to observe statistically significant values. LSTM proved to be the best in terms of accuracy score with model explaining overall 89.98% accuracy in overall snowfall prediction and 90% specificity. Also, it observed the highest precision score of 48% which explains the measure of accuracy with which

the model predicted a snowfall event. However, sensitivity was observed to be the lowest with score of 28.43% seen in table 5.

Table 5: Results of Long Short-Term Memory

Model	Accuracy	Sensitivity	Precision	F1 Score	Specificity
LSTM	89.98	28.43	40	38.66	98

At the end, diagnostic plot was plotted. Which illustrates the performance of the model.by plotting training loss vs test loss. Figure illustrates that the model developed is a good fit example as the model performed well both in training set as well as validation set as seen in figure 15.

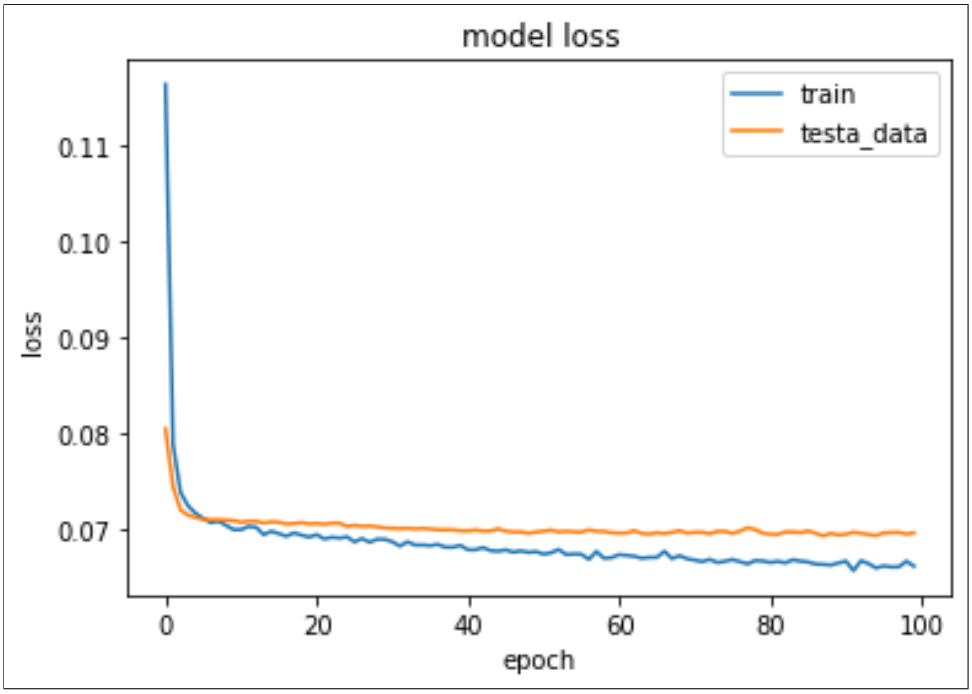

Figure 15: LSTM Diagnostic plot

4.6 Comparison of Developed Models

In this section, all the implemented models Decision tree, Random Forest, XG Boost and Long Short-Term Memory for the prediction of Snowfall in Boston region are discussed and compared based on the taken evaluation metrics Accuracy, Sensitivity (Recall), Specificity, Precision, and F1. From the figure 16, it is clearly evident that in terms of Accuracy obtained by Long Short-Term Memory (LSTM) is greater than all the other developed models. It was successful in predicting the highest number of snowfall events correctly by 89.98% compared to all other models whereas Random forest observed the lowest accuracy of 78.24%.

All the developed models were implemented for the same top 20 important features where the dataset was split in 80-20 training and test set. In this analysis, focus was to achieve the highest sensitivity score which tells us about the event when there was a

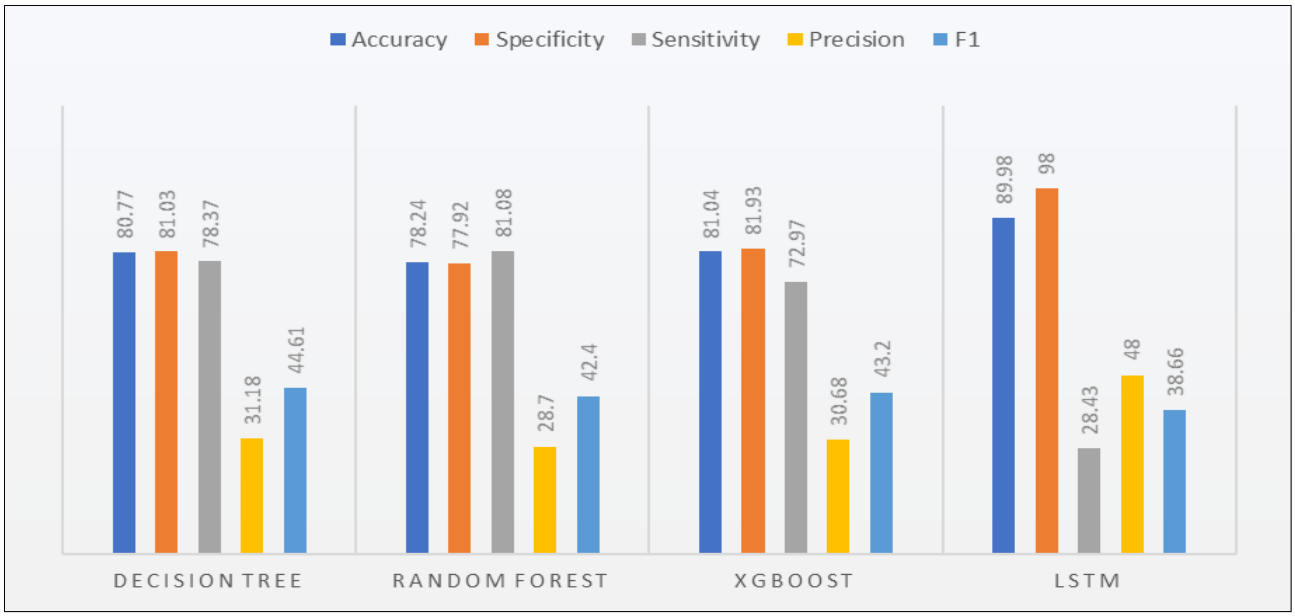

Figure 16: Coefficients Multiple linear regression

snowfall and the model correctly predicted it. For which, interestingly model Random Forest recorded the highest sensitivity score of 81.08% and least was observed by LSTM which results to conclude that Random forest is suitable for this dataset and outperformed others. It was observed that that LSTM achieved with the highest precision of 48% which means whenever it predicted a snowfall event, the accuracy with which it classified it was highest recorded. Similarly, for specificity both the Decision tree and XG Boost recorded a score of 81%. Based on the overall results for every chosen metric XG Boost and Decision Tree performed well in the classification of snowfall prediction

4.7 Conclusion

On the basis of the developed models and the results produced by them this technical report fully answered research question addressed in section 1.2. Also, all the mentioned objectives mentioned in section 1.3 were undertaken. The results obtained from the implementation of these machine learning models will significantly contribute to the body of knowledge in the domain of snowfall prediction.

5 Conclusion and Future Work

This research analysis proposed the implementation of Machine learning algorithms Decision tree, Random Forest, XG Boost and Long Short-Term Memory for which aim was to predict Snowfall in the region of Boston for 10 years of meteorological data which has been successfully achieved by employing machine learning techniques. The models were trained on the climate features of (temperature, humidity, wind, dew point, precipitation, sea level, and visibility) for which daily average, highest and lowest values were recorded. In addition, the results of this research added to the contribution of the knowledge in weather prediction in the domain of Snowfall machine learning industry. This is the first research to explore the applied deep learning LSTM for the classification of Snowfall prediction. The results acquired in this technical report is promising. LSTM

model outperformed other models in terms of accuracy for snowfall prediction with an accuracy of 89.98% but as sensitivity score is aimed to achieve in this analysis Random Forest outperformed every other model. Also, it was concluded that overall Decision Tree and XG Boost performed well comparatively keeping in considerations all the metrics. These results were noted for all the yearly data but more accurate results could have been drawn by taking the meteorological values for only winter months and on a large dataset for better results in terms of deep learning.

Therefore, though the models achieved high performance there is assuredly room for improvement for increased accuracy. We analyzed the models for yearly data where all the summer months were also included which may not have given optimum results as the data was biased. Additionally, a much large dataset should be endorsed so we can only include winter months of the region and implement deep learning as more the training data better the model can be trained.

In future work, more affecting factors responsible for a snowfall event can be included for Snowfall prediction. An extensively large data can betaken for only winter months in order to sufficiently train a deep learning model. Also, the advanced time-series models like GRU and Prophet could be used to for implementation which has shown success with other time-series domain.

References

Aftab, S., Ahmad, M., Hameed, N., Bashir, S., Ali, I. and Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review, *International Journal of Advanced Computer Science and Applications* **9**.

Alnoukari, M. and El Sheikh, A. (2012). *Knowledge Discovery Process Models: From Traditional to Agile Modeling*, pp. 72–100.

Chao, Z., Pu, F., Yin, Y., Han, B. and Chen, X. (2018). Research on real-time local rainfall prediction based on mems sensors, *J. Sensors* **2018**: 6184713:1–6184713:9.

Chatterjee, S., Datta, B., Sen, S., Dey, N. and C. Debnath, N. (2018). Rainfall prediction using hybrid neural network approach, pp. 67–72.

Dubey, A. (2015). K-means based radial basis function neural networks for rainfall prediction, pp. 1–6.

Geetha, A. and Nasira, G. M. (2014a). Data mining for meteorological applications: Decision trees for modeling rainfall prediction, *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–4.

Geetha, A. and Nasira, G. M. (2014b). Data mining for meteorological applications: Decision trees for modeling rainfall prediction, *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–4.

Holmstrom, M. A. and Liu, D. Z. (2016). Machine learning applied to weather forecasting.

Hossain, M., Rekabdar, B., J. Louis, S. and Dascalu, S. (2015). Forecasting the weather of nevada: A deep learning approach, pp. 1–6.

Joshi, J. (2012). Prediction of weather states using hidden markov model.

Kala, A. and Vaidyanathan, S. (2018). Prediction of rainfall using artificial neural network, pp. 339–342.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T. and Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather, *Bulletin of the American Meteorological Society* **98**(10): 2073–2090.
URL: *https://doi.org/10.1175/BAMS-D-16-0123.1*

Mohapatra, S. K., Upadhyay, A. and Gola, C. (2017). Rainfall prediction based on 100 years of meteorological data, *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, pp. 162–166.

Navid, M. and Niloy, N. (2018). Multiple linear regressions for predicting rainfall for bangladesh.

Roebber, P., R. Butt, M., J. Reinke, S. and J. Grafenauer, T. (2007). Real-time forecasting of snowfall using a neural network, *Weather and Forecasting - WEATHER FORECAST* **22**.

Sejian (2016). Impact of climate change on livestock productivity, fao food and nutrition series.

Srinivasan, K., Semwal, G. and Sunil, T. (2013). Statistical-based forecasting of avalanche prediction, *Defence Science Journal* **49**(5): 447–455.
URL: *http://publications.drdo.gov.in/ojs/index.php/dsj/article/view/3859*

Tae Sohn, K., Hyeong Lee, J. and Seuk Cho, Y. (2009). Ternary forecast of heavy snowfall in the honam area, korea, *Advances in Atmospheric Sciences* **26**: 327–332.

Tang, G., Long, D., Behrangi, A., Wang, C. and Hong, Y. (2018). Exploring deep neural networks to retrieve rain and snow in high latitudes using multisensor and reanalysis data, *Water Resources Research* **54**(10): 8253–8278.

Tharun, V., Prakash, R. and Subramanian, R. d. (2018). Prediction of rainfall using data mining techniques, pp. 1507–1512.

Thirumalai, C. S., Sri Harsha, K., Lakshmi Deepak, M. and Chaitanya Krishna, K. (2017). Heuristic prediction of rainfall using machine learning techniques.

Xueting ZHANG, Xuemei LI, L. L. S. Z. Q. Q. (2019). Environmental factors influencing snowfall and snowfall prediction in the tianshan mountains, northwest china, *Journal of Arid Land* **11**(1): 15.
URL: *http://jal.xjegi.com/EN/abstract/article*$_5$*82.shtml*

Zaytar, M. A. and El, C. (2016). Sequence to sequence weather forecasting with long short-term memory recurrent neural networks, *International Journal of Computer Applications* **143**: 7–11.

Zhang, X., Li, X., Li, L., Zhang, S. and Qin, Q. (2019). Environmental factors influencing snowfall and snowfall prediction in the tianshan mountains, northwest china, *Journal of Arid Land* **11**: 15–28.