**KLE Society's**

**KLE Technological University**



**Data Mining and Analysis - 18ECSC301**

**Course Project Report**

**on**

**WEATHER SHIFT PREDICTION**

**Bachelor of Engineering**

**In**

**Computer Science and Engineering**

**Submitted By**

**Team No: 13**

|  |  |  |
| --- | --- | --- |
| **NAME** | **ROLL NO** | **SRN** |
| **Srishti S Eligar** | **137** | **01FE19BCS042** |
| **Ankita A Mane** | **146** | **01FE19BCS052** |
| **Swathi Pardeshi** | **153** | **01FE19BCS060** |
| **Muskan R Havaldar** | **164** | **01FE19BCS007** |

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING**

HUBLI–580031 (India).

Academic year 2021-2

1. Introduction
   1. Problem Statement
   2. Overview of Problem Statement
   3. Description of Project
2. Understanding data
   1. Dataset Description
   2. Attributes Description
   3. Preprocessing
3. Learning models
4. Result
5. Ranking
6. References
7. **Introduction**
   1. **Problem Statement**

The goal is to predict the temperature at a particular latitude/longitude and time, given all available measurements and climate model predictions.

* 1. **Overview of Problem Statement**

Weather Forecasting, A forecast refers to a calculation or an estimation which uses data from previous events, combined with recent trends to come up a future event outcome. Weather Shift Prediction, Prediction is an actual act of indicating that something will happen in the future with or without prior information. The goal is to predict the temperature at a particular latitude/longitude and time, given all available measurements and climate model predictions. Training data is not uniformly distributed across different climate zones. Weather systems are highly non-stationary processes.

* 1. **Description of Project**

Prediction is an actual act of indicating that something will happen in the future with or without prior information. Weather shift prediction is the prediction for long-term alteration of temperature and typical weather patterns in a place. We propose the Shifts Dataset for evaluation of uncertainty estimates and robustness to distributional shift. Weather warnings are important forecasts because they are used to protect life and property. Since outdoor activities are severely curtailed by heavy rain, snow and wind chill, forecasts can be used to plan activities around these events, and to plan ahead and survive them. Weather forecasters use all kinds of tools to achieve this goal.

1. **Understanding the data**

We have 3 types of data as .csv file – train , dev (dev-in and dev-out) and eval.

* 1. **Data description**

1. **Train.csv**

Number of Rows: 3129592

Number of Columns: 129

Size: 5.03 GB

1. **Dev\_in.csv and Dev\_out.csv**

* Dev-in

Number of Rows: 50000

Number of Columns: 129

Size: 81.7 MB

* Dev-out

Number of Rows: 50000

Number of Columns: 129

Size: 80.2 MB

1. **Eval.csv**

Number of Rows: 1137731

Number of Columns: 123

Size: 1.71 GB

* 1. **Attribute description**

**Meta-data**

| Name | Description |
| --- | --- |
| fact-time | Time at a particular instance (it is a 10-digit value representing date and time) |
| fact-latitude | Particular latitude that represents a region. |
| fact-longitude | Particular longitude that represents a region. |
| fact-temperature | Temperature predicted at a particular time. |
| fact-cwsm-class | Canadian Meteorological statute mile. (a unit length=1760 yards/1609.344 meters) |
| climate | Climatic conditions like tropical, mild temperature, dry and snow. |

### Other attributes

| Name | Description |
| --- | --- |
| CMC | Canadian Meteorological Center(CMC) model , 56 attributes |
| WRF | Weather Research and Forecasting (WRF) model , 13 attributes. The Weather Research and Forecasting (WRF) Model is a numerical weather prediction system designed for both atmospheric research and operational forecasting applications. |
| GRF | Global Forecast system predictions(GFS) , 50 attributes. The Global Forecast System (GFS) is a global numerical weather prediction system. |
| Topography-bathymetry | Topography describes physical features of land in a specific region. Bathymetry refers to the study of the ocean's depth relative to sea-level. |
| Sun-elevation | Sun Elevation is an angular coordinate system for locating the sun's positions in the sky. |
| Climate-temperature | Temperature recorded at a particular time. |
| Climate-pressure | Pressure recorded at a particular time. |

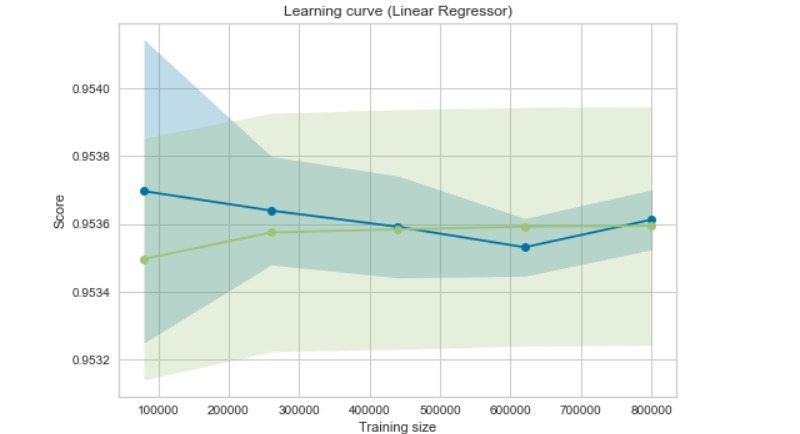
* 1. **Preprocessing**

Initially we removed rows containing blank cell are removed. Columns that do not affect the modeling process majorly are removed. 32% of dataset used (10,00,000 rows x 113 columns).

1. **Learning models**
   1. **LINEAR REGRESSION MODEL**

Multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

* The independent variable is temperature.
* The dependent variables are all other columns other than meta data and temperature.
* ***RMSE: 2.168***

******

* 1. **ELASTIC NET MODEL**

Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions.

**PRE-PROCESSING -**

* Removing rows with missing values
* Removing columns
* Columns removed: gfs\_temperature\_(15000….97500)

Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions.

* ***RMSE: 2.178***
  1. **XG BOOST**

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

**PRE PROCESSING1 -**

* Removing rows with missing values
* Removing columns
* *Columns removed: gfs\_temperature\_(15000….97500)*
* Parameters:
* colsample\_bytree = 0.4, gamma=0, learning\_rate=0.3, max\_depth=6, min\_child\_weight=1, n\_estimators=100, reg\_alpha=0.75, reg\_lambda=0.45, subsample=0.6, seed=42.
* ***RMSE: 1.978***

**PRE PROCESSING2 -**

* Removed metadata column
* Filled missing cells with 0
* ***RMSE: 4.49***

**PRE PROCESSING3 -**

* Feature selection using F-score
* ***RMSE: 1.948***
  1. **CATBOOST REGRESSOR**

CatBoost builds upon the theory of decision trees and gradient boosting.It introduces ordered boosting as a better gradient boosting algorithm

**PRE PROCESSING -**

* Removed metadata column
* Filled missing cells with 0
* Parameters: Iteration=2000, Learning rate = 0.04
* ***RMSE: 1.9745(dev\_in and dev\_out)***
* ***RMSE: 2.717(dev\_in and dev\_out)***

1. **Result**

|  |  |
| --- | --- |
| **MODEL** | **RMSE value** |
| LINEAR REGRESSION | 2.168 |
| ELASTIC NET | 2.178 |
| XGBoost | 1.978 and 4.49 and 1.948 |
| CatBoost | 1.9745 and 2.717 |

1. **Ranking**



1. **References**

* Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks ([Andrey Malinin](https://arxiv.org/search/cs?searchtype=author&query=Malinin%2C+A), [Neil Band](https://arxiv.org/search/cs?searchtype=author&query=Band%2C+N), [Ganshin](https://arxiv.org/search/cs?searchtype=author&query=Ganshin), [Alexander](https://arxiv.org/search/cs?searchtype=author&query=Alexander), [German Chesnokov](https://arxiv.org/search/cs?searchtype=author&query=Chesnokov%2C+G), [Yarin Gal](https://arxiv.org/search/cs?searchtype=author&query=Gal%2C+Y), [Mark J. F. Gales](https://arxiv.org/search/cs?searchtype=author&query=Gales%2C+M+J+F), [Alexey Noskov](https://arxiv.org/search/cs?searchtype=author&query=Noskov%2C+A), [Andrey Ploskonosov](https://arxiv.org/search/cs?searchtype=author&query=Ploskonosov%2C+A), [Liudmila Prokhorenkova](https://arxiv.org/search/cs?searchtype=author&query=Prokhorenkova%2C+L), [Ivan Provilkov](https://arxiv.org/search/cs?searchtype=author&query=Provilkov%2C+I), [Vatsal Raina](https://arxiv.org/search/cs?searchtype=author&query=Raina%2C+V), [Vyas Raina](https://arxiv.org/search/cs?searchtype=author&query=Raina%2C+V), [Roginskiy](https://arxiv.org/search/cs?searchtype=author&query=Roginskiy), [Denis](https://arxiv.org/search/cs?searchtype=author&query=Denis), [Mariya Shmatova](https://arxiv.org/search/cs?searchtype=author&query=Shmatova%2C+M), [Panos Tigas](https://arxiv.org/search/cs?searchtype=author&query=Tigas%2C+P), [Boris Yangel](https://arxiv.org/search/cs?searchtype=author&query=Yangel%2C+B))
* Weather Forecasting Model using Artificial Neural Network ([Kumar Abhishek, M.P.Singh, Saswata Ghosh, Abhishek Anand](https://www.sciencedirect.com/science/article/pii/S221201731200326X))
* An ANN Model Trained on Regional Data in the Prediction of Particular Weather Conditions