SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Singapore University of Technology & Design (SUTD)

40.011 Data and Business Analytics

Executive Summary

Prediction of Lighting Occurrence at Airport

(ASI, Vaisala)

Team 17

24/4/2023

Name	Student ID
Zaina Aafreen d/o Ziyavuddeen	1006145
Tan Jun Onn	1006234
Marcus Tan Zong Wei	1006334
Langarkande Rishita Sanjeev	1006031

Advisor: Dai Gengling

Executive Summary

Accurate and effective prediction of the occurrence of lightning at airports has been a significant issue as lightning strikes pose a serious safety risk to airport personnel and can cause disruptions to operational procedures. This is especially an issue at Don Mueang International Airport in Bangkok as the airport is situated in an area that experiences frequent thunderstorms and lightning strikes. Current Lightning Prediction systems available are faced with the challenge of reducing false alarms while accurately predicting lightning strikes under rapidly changing weather conditions at low cost.

As such, our team has been tasked with predicting the probability of a lightning strike in the region around Don Mueang International Airport where our prediction should be based on lightning strikes in neighboring regions in the past 60 minutes. For our predictive analysis, we had access to data from Aviation Studies Institute (ASI) and Vaisala. Vaisala is a company that specialises in weather prediction technologies like lightning detection and prediction systems while ASI is a research center that hopes to enhance aviation safety and operational efficiency at airports.

For simplicity, we focused on 756.25km^2 area of the airport region by creating a 5x5 grid of 5.5km squares where we segmented each grid square by their longitude and latitude boundaries in SQL. After our preliminary analysis, using R, we discovered the occurrence of lightning in 7 surrounding grid squares at time t-1 around the airport to be significant factors that contribute to the prediction of lightning at the airport at time t. Then, we optimized the probability threshold, p=0.03 for more accurate prediction of lightning occurrence. Upon comparison of Accuracy, Sensitivity and Specificity of models, we determined that the 30-minute model is more accurate than the 60-minute model by at least 17%. In efforts to further improve our model, we sourced and included weather data such as Temperature, Dew Point, Humidity, Wind Gust, Wind Speed and Atmospheric Pressure of the area into our analysis. However, we found that the weather data was insignificant at 90% certainty. Thus, we discarded it. Finally, we obtained our optimal 30-minute model with a 98.6% accuracy rate.

Our team recognises that the strength of using our model is that one can reuse the model and translate the predictability to smaller airports around the ASEAN region easily. However, moving forward, we feel that considering data about the wind direction of each of the grid squares, (not relative to each grid square) would improve the model as wind direction is a significant factor that affects occurrence of lightning. Moreover, we feel that further analysis using 15-minute intervals instead would yield greater accuracy and produce a more reliable model that would be able to predict the occurrence of lightning under rapidly changing weather conditions.

In conclusion, our team understands that the development of a predictive model for lightning strikes is crucial for ensuring the safety of airport personnel and maintaining operational efficiency. We recognise our predictive model for the Don Mueang International Airport in Bangkok that has achieved a high accuracy rate of 98.6% would serve as a valuable tool for airport authorities to make informed decision to address safety concerns and optimize operational procedures in aviation. We hope that our research inspires further research in this area.

Company Introduction

For our project, we received 3 months of lightning strike records and their intensity in Southeast Asia from Vaisala and The Aviation Studies Institute (ASI). Vaisala is a Finnish firm that specialises in environmental and industrial measurement. Vaisala aims to assist customers in making educated decisions based on precise and trustworthy data. ASI is a research center in Singapore University of Technology and Design (SUTD) jointly established with the Civil Aviation Authority of Singapore (CAAS). ASI aims to perform cutting-edge research in the aviation field to provide practical solutions to challenges faced in the aviation industry specifically in the Asia-Pacific region.

Problem Definition

The occurrence of lightning poses significant risk and disruptions to airport services and ground staff. It is crucial to warn and remove ground staff from accident prone areas in the event of lightning occurrence to ensure safety and operational efficiency. This is especially an issue of importance at Don Mueang International Airport as it is in a region of Thailand where frequent thunderstorms and lightning strikes occur. As such, our team was tasked **to predict the probability of a lightning strike in the region around Don Mueang International Airport (DMK) at Bangkok, Thailand** based on lightning strikes in neighboring regions in the past 60 minutes.

Methodology

Step 1: Determining Parameters using SQL

We started off our analysis by determining the exact parameters of the area that we were going to examine. Analysing the area between latitudes 13.7883 and 14.0383 and longitudes 100.4792 and 100.7292, we marked out the 756.25km^2 area region using a query (Figure A1). We created a 5x5 grid of 5.5km squares (Figure C1) and labelled the cells from 1 to 25 where cell 13 encompasses the entire area of **Don Mueang International Airport, DMK** (Figure A2). We accomplished this by creating the "CellExtentsThailand" table containing manually input grid boundaries for the latitudes and longitudes (ce.toplat, ce.botlat, ce.leftlong, and ce.rightlong) of each grid square (cellindex). Similarly, the data for the "TimeSegments" table was manually imported using an Excel-generated file (Figures A3) where lightning occurrence was filtered based on time and date.

Step 2: QGIS Data Visualisation

Using our filtered lightning data, elevation data and Thailand Shapefiles, we did a simple preliminary analysis of DMK via QGIS (Figure C2) which helped identify potential visually identifiable geographical relationships with the lightning strikes. Using the Thailand elevation shapefile, we determined that the entire Bangkok Metropolitan Area was situated on flatland and that elevation posed no correlation to the amount of lightning strikes in the area.

Step 3: Binary, Filtering, Time Segmenting and Factoring using SQLite

The query, Q020StrikeCountByCellTimeSegmentThailand (Figure A4) was used to aggregate the strike counts by time segment, cell index and date. This was crucial as each lightning occurrence had a unique "rowid" which could later be used during binary factoring to numerically classify the time of each row as it is detrimental for the prediction model. The next view, “Q030TimeSegmentsCellExtentsThailand” (Figure A5) combined the “TimeSegments”, “CellExtentsThailand” tables which were manually constructed, to use JOIN clauses easily between future views. Another query (Figure A6) combined Q030 and Q020, each with a distinct "rowid" and "cellindex", to provide a timeline view (Figure A7) that depicts the occurrence of lightning at each "cellindex" at every segmented 30-interval via the column “StrikeCount”. Furthermore, to verify that the lightning count was binary for logistics regression, the query (Figure A8) comprised of an "IF-ELSE" statement with the input of 1 in "BinaryCount" if lightning was present and 0 otherwise. Then we used the query “Q060BinaryFactorStrikeCountsThailand” (Figure A9) to form a matrix consisting of binary values for individual cells. However, this presented repetition where every rowid had 25 copies for each cell row and column. To solve this, we used the query “Q070BinaryFactorGroupStrikeCountsThailand” (Figure A11) to transpose each row of “TimeSegment” with the same “rowid” to columns labelled Cell 1 to Cell 25 within 1 unique rowid. To avoid repetition within the cells, we used the max() function to ensure that we obtain the maximum value for each transposed column where only 1 value per cell is returned (Figure A11). Then, we exported the final view and its data (Figure C3) into “lightningstrikedata.csv”.

Step 4: Web scraping weather data on Python

To examine the significance of weather variables in predicting lightning strikes in DMK, it was necessary to web scrape weather data. Selenium and CSV were the only 2 libraries used for web scraping weather data on www.wunderground.com. Selenium iterated through significant dates in the URL as in Figure B1. The chrome driver searched for the weather data values using the function find_elements(BY.XPath) where XPath is a HTML classifier on the website. It utilized a simple nested FOR loop to iterate within each column and then the row that follows in the table that contained the values. The data obtained was then appended into an empty list and written into a CSV file using the “csv.writer()” function (Figure B2). On top of that, we also had to test for linear independence of weather data through a correlation plot (Figure C4). However, no conclusion could be made on the independence on the data.

Step 5: Cleaning and preprocessing in Excel

To balance the number of rows, we first duplicated the cell 13 column, deleted its first cell, and the last cell of the other columns. We then renamed the duplicated column to “NextStrike” to serve as our predictor value for our logistics regression. Second, we added the weather data to their respective hourly "TimeSegments" in the "lightningstrikedata.csv". Now, we have obtained data that was ready for logistics regression using a Generalized Linear Model, GLM (Figure C5).

Step 6: Formulating prediction model using R.

We decided to use Logistics Regression to construct our prediction model for lightning occurrence in DMK (Figure B3) by applying the "Generalized Linear Model” function, glm() in R. This allowed us to prune grid cells and weather conditions from the prediction model that were insignificant[1] and to obtain the optimal predictor coefficients in the form of β_i (Figure B4). As attached, the β_i takes values under “estimates” for each predictor binary variable, x_i.

Step 7: Optimizing prediction threshold, p using R.

After obtaining the prediction model, we used the predict() function to calculate the probability of the next lightning strike occurring at Cell 13 (Figure B5). Given each probability for each “TimeSegment”, we construct the confusion matrix (Figure B5) using the table() function assuming p=0.5 as a temporary variable as a preliminary analysis. To further

[1] Cells that did not exceed 95% significance.

optimize p[2], we constructed an ROC[3] Curve using the performance() function (Figure B6) where "tpr" and "fpr" represents the true positive rate for the y-axis and false positive rate for the x-axis. The graph output cycles a set of p values from 0 to 1 and plots the relationship of the rates accordingly. We then analysed the curve plotted to pick the best p that maximized the ratio of true positive rate to false positives rate (Figure C6). To confirm the optimal p* that we found, we tested our model by deriving their $Accuracy = \frac{TP^4+TN^5}{TP+TN+FN^6+FP^7}$, $Specificity = \frac{TN}{TN+FN}$ and $Sensitivity = \frac{TP}{TP+FP}$ from the confusion matrix (Figure CX) which gives us a numerical understanding for the p* value found from the ROC curve.

Step 8: Data Visualisation in Tableau

Finally, to visualize the individual lightning strikes in the area, we used the "Q040ThaiVizTest" query (Figure A12) to combine Q030 and Q020 which contains the aggregated lightning strikes per cellid and their time segmentation. We exported the data as a CSV in order for it to be read in Tableau. Next, we filtered the data based on "Segmented Index" and "Date". After which, we added new boundary marks with the cell index attribute and their corresponding longitude and latitude (Figure C7) to create the grid box visual.

[2] p= the threshold probability where when the logistic regression model outputs a value exceeding p, it will predict a lightning strike in binary as 1

[3] Receiver Operating Characteristic, ROC plots a graph of True Positive rate against False Positive rate to find the optimum p* value.

[4] TP (True Positive): The number of lightning strikes correctly predicted to occur.

[5] TN (True Negative): The number of lightning strikes correctly predicted not to occur.

[6] FN (False Negative): The number of lightning strike incorrectly predicted not to occur.

[7] FP (False Positive): The number of lightning strike incorrectly predicted to occur.

Results

Analysis of model based on significance of variables and AIC Value

From step 6 of the methodology, we obtained formulations of two types of models, one model that predicted lightning strikes in 60-minute intervals and another that predicted it in 30-minute intervals. We decided to derive these two different types of models to determine their accuracy at a later stage of our analysis.

As illustrated in Figures C9 and C10, for the formulations of the 60-minute and 30-minute model, we derived the best model for each type by taking the Akaike Information Criterion (AIC) values of each formulation and the significance of the variables into account. Since we pruned out variables that were not at least 95 % significant, our final formulations for both types of the model do not have the lowest AIC value. The AIC value measures the goodness of a logistic regression model while including a penalty that is an increasing function of the number of estimated parameters to account for overfitting. Hence, a lower AIC value is preferred. Although, the AIC of our final formulation of both types of models is not the lowest, the disparity between the lowest AIC and our final formulations' AIC are small enough to be overlooked.

The final prediction model for both types follow a Binary Logistic Regression equation $Probability = \frac{1}{1+e^{-f(x)}}$ where $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$. Each binary coefficient, β_i corresponds to binary variables, x_i for occurrence of lightning for selected significant cells where the variable value 1 indicates occurrence of lightning while the variable value 0 indicates the absence of lightning. As such we derived our final Binary Logistic Regression equation for the 2 models. (Figure C11).

Analysis of model based on Accuracy, Specificity and Sensitivity

From step 7 of the methodology, we determined the probability threshold, p to be 0.03 after analysing the ROC curve in Figure C6. Hence, the model will only predict that lightning may occur if its probability is greater than 0.03; otherwise, the model will predict that no lightning strikes may occur within a given interval. Using this we constructed a confusion matrix (Figure CX) with 2 dimensions of "Actual" against "Predicted with 4 classes of values: True Negative (TN) True Positive (TP), False Negative (FN) and False Positive (FP).

Using the confusion matrices in Figure C12, we derived the Accuracy, Sensitivity and Specificity (ASS) of both models. Upon comparison of ASS in Figure C13, we easily determined that the 30-minute model is more optimal as it has higher ASS values. Thus, we selected the 30-minute model and illustrated the significant variables of the model in Figure C8 for clear understanding.

Testing our model with other ASEAN airports

We conducted additional testing of our final model with optimal p=0.03 by applying our model to other lighting strike datasets like Juanda International Airport (JIA) and Kuala Lumpur International Airport (KLIA). We then constructed confusion matrices for each of the Airports (Figure C14) to derive their ASS values. Upon comparison of ASS values in Figure C15, we conclude that our model is applicable to DMK, JIA and KLIA as all three airports have high Accuracy and Specificity values of above 98%. This could be the result of the presence of high number of instances where lightning strikes do not occur which skews the results that predicted no lightning accurately (negative cases).

We also observed that the sensitivities of all 3 airports are comparatively lower than accuracy and specificity values with the best sensitivity in DMK and worst sensitivity in JIA. This might be because there were fewer lightning incidents in the dataset for JIA than at DMK and

KLIA. When compared to the monsoon seasons in Bangkok (July to October) and Kuala Lumpur (October to January), Surabaya's monsoon season (December to March) has the least overlap with the lightning data gathered (September to December). As a result, Surabaya's lightning data may be statistically less valuable to our model than that from the other two places because it only covers a month.

Weather data Analysis

From step 4 of the methodology, we obtained weather data such as temperature, dew point, humidity, wind gust, wind speed and atmospheric pressure. However, after we ran the additional variables through R, we discovered that the weather variables were not of significance. Thus, the weather variables were pruned from the model. After further analysis, we feel that the insignificance of the weather data could be attributed to the data that was sourced. The data we sourced was obtained from a single weather tower that spanned over all the 25 grid cells. Thus, resulting in constant weather data over all the cells. Hence, we hypothesize that weather data could be of significance to the model if we are able to obtain weather variable data specific to each individual cell.

Evaluation

Limitation	**Improvement**
We assumed that there were no weather or seasonal changes over the year. Hence, our model might not be applicable or specific to different seasons of year	Collect year-round data to split the year into seasons. This would generate season specific models for more accurate prediction.
Data collected online about wind direction was unsuitable for use as wind direction data was not with respect to cell 13. Hence, we were unable to include it our analysis	Create a vector model to account for the wind direction with respect to cell 13 or set up weather tracking device at cell 13 to physically collect data on wind direction.
Our model only predicts the occurrence of lightning but not the frequency of lightning strikes.	Reduce the size of the grid squares (e.g., 0.1km by 0.1km) such that the location of the lightning strike can be obtained more precisely. By adding up the number of lightning strikes in the smaller grid, we can obtain the total number of lightning strikes in the larger grid (5.5km by 5.5km).

References

1. Long, T., & Miller, I. (2022, June 13). *Impact of Lightning Strikes on Airport Facility and Ground Operations*. View of impact of lightning strikes on airport facility and Ground Operations. Retrieved from, https://ojs.library.okstate.edu/osu/index.php/CARI/article/view/8533/7785
2. *Bangkok, Thailand weather conditionsstar_ratehome*. Weather Underground. (n.d.). Retrieved April 23, 2023, from https://www.wunderground.com/hourly/VTBD
3. Ying Xu. (2023, February). *Logistic Regression*. *Engineering Systems Design*

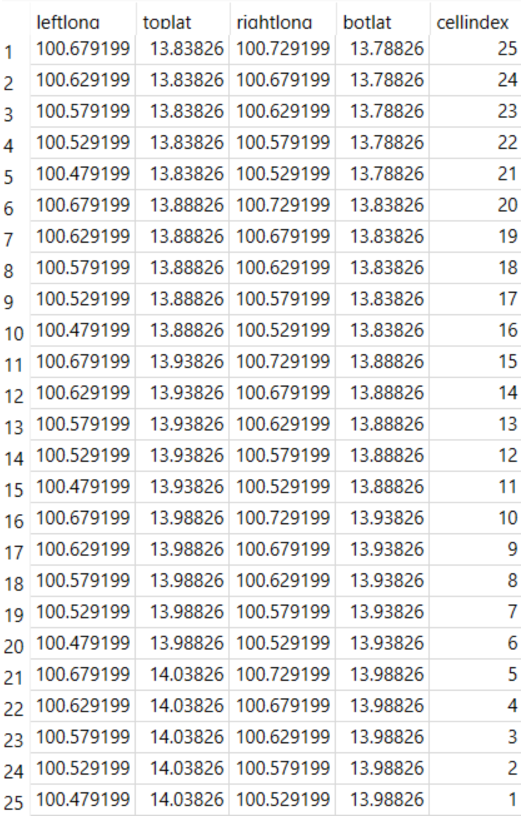
Appendix

Appendix A – SQL

Figure A1 (Analyzing data within the area)

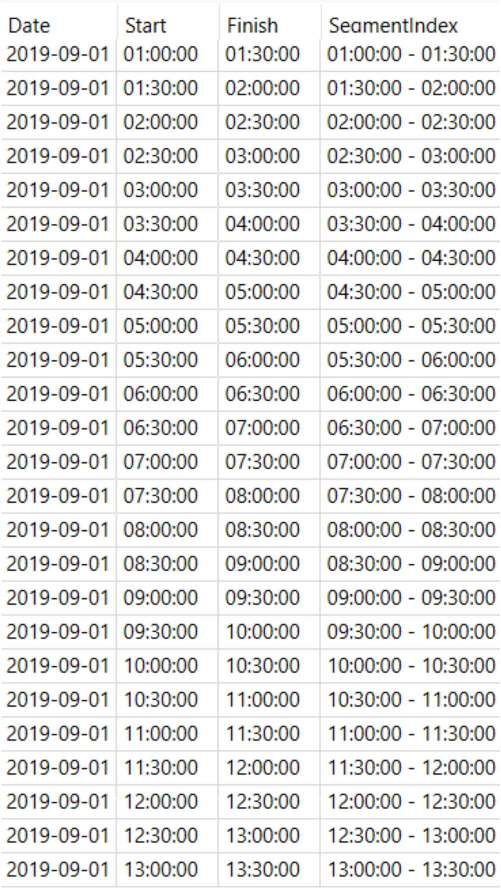
```
SELECT *
  FROM Lightning
 WHERE "Latitude" <= (13.91326+ 0.124) AND
       "Latitude" >= (13.91326-0.124) AND
       "Longitude" <= (100.604199 + 0.124) AND
       "Longitude" >= (100.604199 - 0.124)
```

Figure A2 (Creating the boundaries of each cell)



	leftlong	toplat	rightlong	botlat	cellindex
1	100.679199	13.83826	100.729199	13.78826	25
2	100.629199	13.83826	100.679199	13.78826	24
3	100.579199	13.83826	100.629199	13.78826	23
4	100.529199	13.83826	100.579199	13.78826	22
5	100.479199	13.83826	100.529199	13.78826	21
6	100.679199	13.88826	100.729199	13.83826	20
7	100.629199	13.88826	100.679199	13.83826	19
8	100.579199	13.88826	100.629199	13.83826	18
9	100.529199	13.88826	100.579199	13.83826	17
10	100.479199	13.88826	100.529199	13.83826	16
11	100.679199	13.93826	100.729199	13.88826	15
12	100.629199	13.93826	100.679199	13.88826	14
13	100.579199	13.93826	100.629199	13.88826	13
14	100.529199	13.93826	100.579199	13.88826	12
15	100.479199	13.93826	100.529199	13.88826	11
16	100.679199	13.98826	100.729199	13.93826	10
17	100.629199	13.98826	100.679199	13.93826	9
18	100.579199	13.98826	100.629199	13.93826	8
19	100.529199	13.98826	100.579199	13.93826	7
20	100.479199	13.98826	100.529199	13.93826	6
21	100.679199	14.03826	100.729199	13.98826	5
22	100.629199	14.03826	100.679199	13.98826	4
23	100.579199	14.03826	100.629199	13.98826	3
24	100.529199	14.03826	100.579199	13.98826	2
25	100.479199	14.03826	100.529199	13.98826	1

Figure A3 (Time Segment data for every half hour)



Date	Start	Finish	SegmentIndex
2019-09-01	01:00:00	01:30:00	01:00:00 - 01:30:00
2019-09-01	01:30:00	02:00:00	01:30:00 - 02:00:00
2019-09-01	02:00:00	02:30:00	02:00:00 - 02:30:00
2019-09-01	02:30:00	03:00:00	02:30:00 - 03:00:00
2019-09-01	03:00:00	03:30:00	03:00:00 - 03:30:00
2019-09-01	03:30:00	04:00:00	03:30:00 - 04:00:00
2019-09-01	04:00:00	04:30:00	04:00:00 - 04:30:00
2019-09-01	04:30:00	05:00:00	04:30:00 - 05:00:00
2019-09-01	05:00:00	05:30:00	05:00:00 - 05:30:00
2019-09-01	05:30:00	06:00:00	05:30:00 - 06:00:00
2019-09-01	06:00:00	06:30:00	06:00:00 - 06:30:00
2019-09-01	06:30:00	07:00:00	06:30:00 - 07:00:00
2019-09-01	07:00:00	07:30:00	07:00:00 - 07:30:00
2019-09-01	07:30:00	08:00:00	07:30:00 - 08:00:00
2019-09-01	08:00:00	08:30:00	08:00:00 - 08:30:00
2019-09-01	08:30:00	09:00:00	08:30:00 - 09:00:00
2019-09-01	09:00:00	09:30:00	09:00:00 - 09:30:00
2019-09-01	09:30:00	10:00:00	09:30:00 - 10:00:00
2019-09-01	10:00:00	10:30:00	10:00:00 - 10:30:00
2019-09-01	10:30:00	11:00:00	10:30:00 - 11:00:00
2019-09-01	11:00:00	11:30:00	11:00:00 - 11:30:00
2019-09-01	11:30:00	12:00:00	11:30:00 - 12:00:00
2019-09-01	12:00:00	12:30:00	12:00:00 - 12:30:00
2019-09-01	12:30:00	13:00:00	12:30:00 - 13:00:00
2019-09-01	13:00:00	13:30:00	13:00:00 - 13:30:00

Figure A4 (Query Q020StrikeCountByCellTimeSegmentThailand)

```
select Q010.rowid,Q010.segmentindex,Q010.cellindex,Q010.Date,
COUNT(Q010.PeakCurrent) as StrikeCount
from Q010StrikesByCellByTimeSegmentThailand as Q010
group by Q010.segmentindex,Q010.cellindex,Q010.Date
order by Q010.Date ASC
```

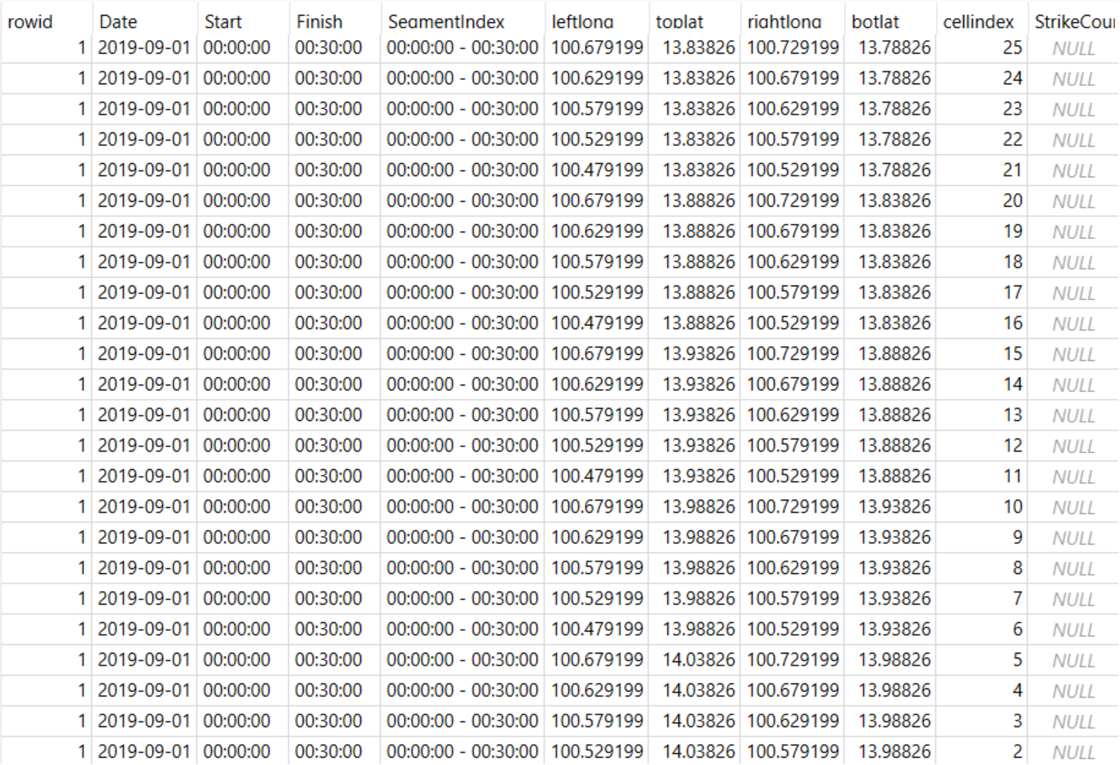
Figure A5 (Query Q030TimeSegmentsCellExtentsThailand)

```
select ts.rowid,ts.*,ce.* from TimeSegments as ts,
CellExtentsThailand as ce
```

Figure A6 (Query that combines queries from Figure A4 & A5)

```
select Q030.*,Q020.StrikeCount from
Q030AllTimeSegmentsCellExtentsThailand as Q030 LEFT JOIN
Q020StrikeCountByCellByTimeSegmentThailand as Q020 on
Q030.rowid=Q020.rowid and Q030.cellindex=Q020.cellindex
```

Figure A7 (Resulting Data Table shown from Figure A6)



rowid	Date	Start	Finish	SegmentIndex	leftlong	toplat	rightlong	botlat	cellindex	StrikeCou
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.679199	13.83826	100.729199	13.78826	25	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.629199	13.83826	100.679199	13.78826	24	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.579199	13.83826	100.629199	13.78826	23	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.529199	13.83826	100.579199	13.78826	22	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.479199	13.83826	100.529199	13.78826	21	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.679199	13.88826	100.729199	13.83826	20	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.629199	13.88826	100.679199	13.83826	19	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.579199	13.88826	100.629199	13.83826	18	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.529199	13.88826	100.579199	13.83826	17	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.479199	13.88826	100.529199	13.83826	16	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.679199	13.93826	100.729199	13.88826	15	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.629199	13.93826	100.679199	13.88826	14	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.579199	13.93826	100.629199	13.88826	13	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.529199	13.93826	100.579199	13.88826	12	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.479199	13.93826	100.529199	13.88826	11	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.679199	13.98826	100.729199	13.93826	10	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.629199	13.98826	100.679199	13.93826	9	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.579199	13.98826	100.629199	13.93826	8	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.529199	13.98826	100.579199	13.93826	7	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.479199	13.98826	100.529199	13.93826	6	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.679199	14.03826	100.729199	13.98826	5	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.629199	14.03826	100.679199	13.98826	4	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.579199	14.03826	100.629199	13.98826	3	NULL
1	2019-09-01	00:00:00	00:30:00	00:00:00 - 00:30:00	100.529199	14.03826	100.579199	13.98826	2	NULL

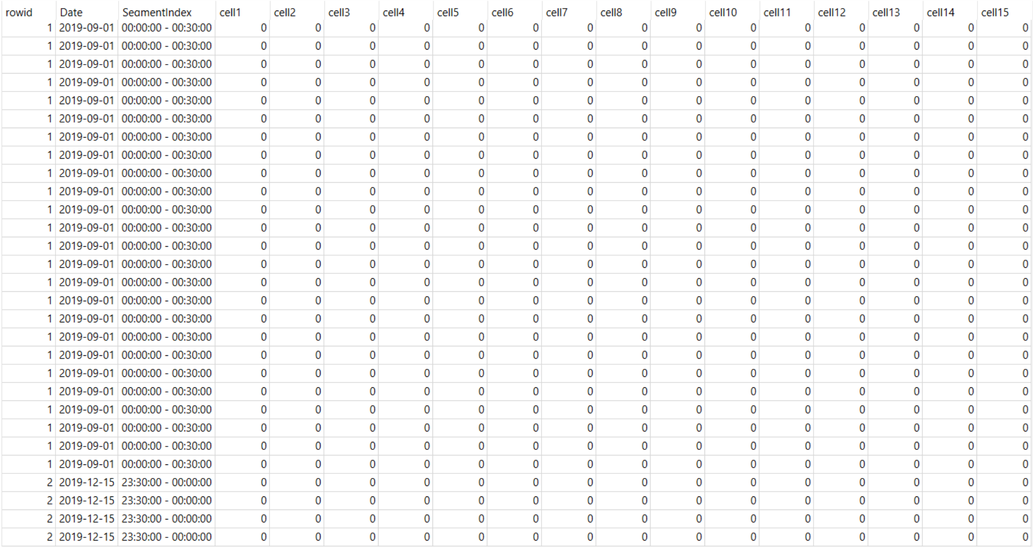
Figure A8 (Query to turn StrikeCount into a Binary Variable)

```
Select Q040.rowid, Q040.Date, Q040.SegmentIndex, Q040.cellindex,
Q040.StrikeCount,
    Case when Q040.StrikeCount is not null
    then 1
    else 0
    End as BinaryCount
from Q040StrikeCountsThailand as Q040
```

Figure A9 (Query Q060BinaryFactorStrikeCountsThailand)

```
Select Q050.rowid,Q050.Date,Q050.SegmentIndex,
    iif(cellindex=1 and BinaryCount=1,1,0) as cell1,
    iif(cellindex=2 and BinaryCount=1,1,0) as cell2,
    iif(cellindex=3 and BinaryCount=1,1,0) as cell3,
    iif(cellindex=4 and BinaryCount=1,1,0) as cell4,
    iif(cellindex=5 and BinaryCount=1,1,0) as cell5,
    iif(cellindex=6 and BinaryCount=1,1,0) as cell6,
    iif(cellindex=7 and BinaryCount=1,1,0) as cell7,
    iif(cellindex=8 and BinaryCount=1,1,0) as cell8,
    iif(cellindex=9 and BinaryCount=1,1,0) as cell9,
    iif(cellindex=10 and BinaryCount=1,1,0) as cell10,
    iif(cellindex=11 and BinaryCount=1,1,0) as cell11,
    iif(cellindex=12 and BinaryCount=1,1,0) as cell12,
    iif(cellindex=13 and BinaryCount=1,1,0) as cell13,
    iif(cellindex=14 and BinaryCount=1,1,0) as cell14,
    iif(cellindex=15 and BinaryCount=1,1,0) as cell15,
    iif(cellindex=16 and BinaryCount=1,1,0) as cell16,
    iif(cellindex=17 and BinaryCount=1,1,0) as cell17,
    iif(cellindex=18 and BinaryCount=1,1,0) as cell18,
    iif(cellindex=19 and BinaryCount=1,1,0) as cell19,
    iif(cellindex=20 and BinaryCount=1,1,0) as cell20,
    iif(cellindex=21 and BinaryCount=1,1,0) as cell21,
    iif(cellindex=22 and BinaryCount=1,1,0) as cell22,
    iif(cellindex=23 and BinaryCount=1,1,0) as cell23,
    iif(cellindex=24 and BinaryCount=1,1,0) as cell24,
    iif(cellindex=25 and BinaryCount=1,1,0) as cell25
from Q050BinaryStrikeCountsThailand as Q050
```

Figure A10 (Resulting Data Table shown from Figure A9)



rowid	Date	SegmentIndex	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10	cell11	cell12	cell13	cell14	cell15
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2019-09-01	00:00:00 - 00:30:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2019-12-15	23:30:00 - 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2019-12-15	23:30:00 - 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2019-12-15	23:30:00 - 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2019-12-15	23:30:00 - 00:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure A11 (Query Q070BinaryFactorGroupStrikeCountsThailand)

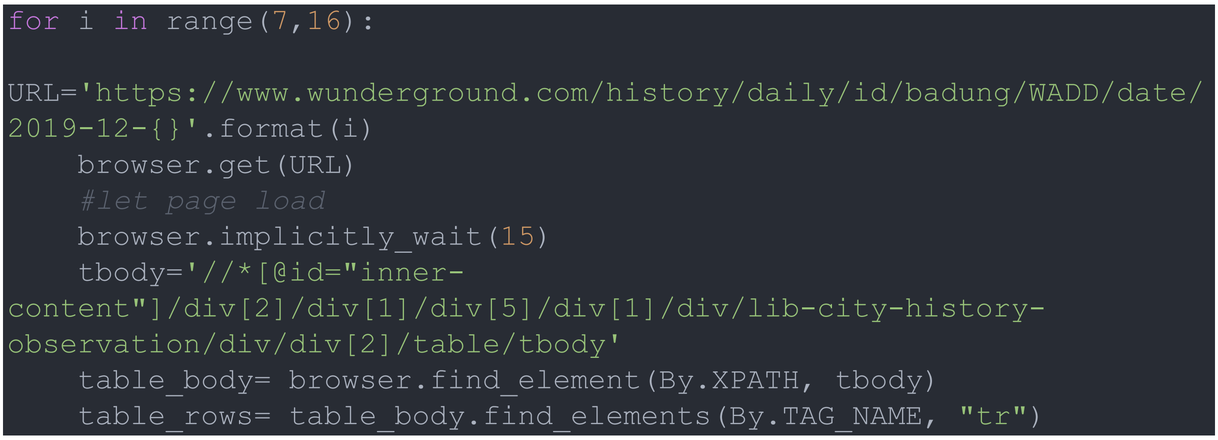
```
select Q060.rowid as TimeSegment,
max(Q060.cell1)as Cell1,max(Q060.cell2)as Cell2,max(Q060.cell3)as
Cell3,max(Q060.cell4)as Cell4,max(Q060.cell5)as
Cell5,max(Q060.cell6)as Cell6,max(Q060.cell7)as
Cell7,max(Q060.cell8)as Cell8,max(Q060.cell9)as
Cell9,max(Q060.cell10)as Cell10,
max(Q060.cell11) as Cell11,max(Q060.cell12)as
Cell12,max(Q060.cell13)as Cell13,max(Q060.cell14)as
Cell14,max(Q060.cell15)as Cell15,max(Q060.cell16)as
Cell16,max(Q060.cell17)as Cell17,max(Q060.cell18)as
Cell18,max(Q060.cell19)as Cell19,max(Q060.cell20)as Cell20,
max(Q060.cell21)as Cell21,max(Q060.cell22)as
Cell22,max(Q060.cell23)as Cell23,max(Q060.cell24)as
Cell24,max(Q060.cell25)as Cell25

 from Q060BinaryFactorStrikeCountsThailand as Q060
 group by Q060.rowid
```

Figure A12 (Query Q040ThaiVizTest)

```
select Q030.*,Q020.StrikeCount,Q020.Latitude,Q020.Longitude from
Q030AllTimeSegmentsCellExtentsThailand as Q030 LEFT JOIN
Q020ThaiVizTest as Q020 on Q030.rowid=Q020.rowid and
Q030.cellindex=Q020.cellindex
```

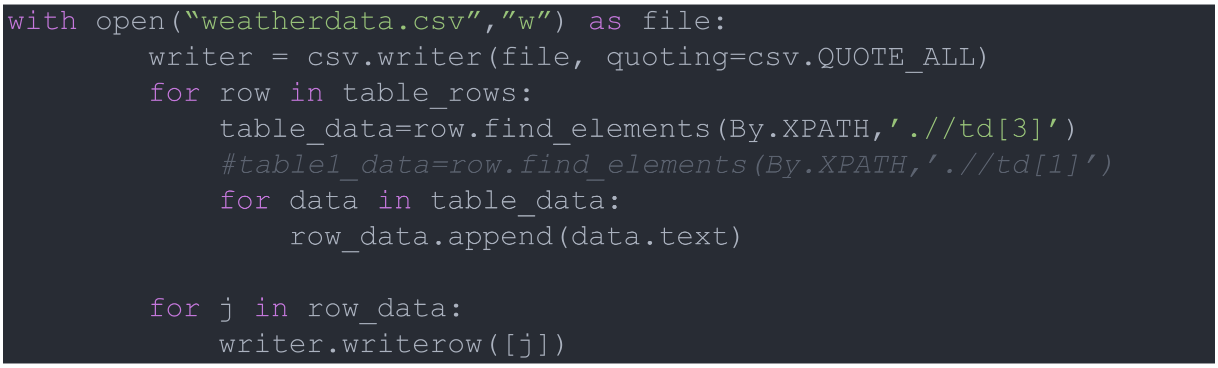
Appendix B – Python & R

Figure B1 (Web Scraping using Selenium)



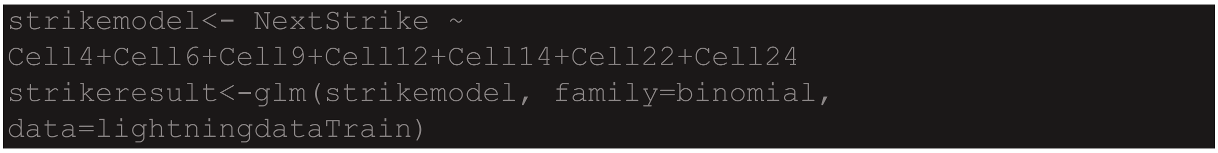
```
for i in range(7,16):

URL='https://www.wunderground.com/history/daily/id/badung/WADD/date/
2019-12-{}'.format(i)
    browser.get(URL)
    #let page load
    browser.implicitly_wait(15)
    tbody='//*[@id="inner-
content"]/div[2]/div[1]/div[5]/div[1]/div/lib-city-history-
observation/div/div[2]/table/tbody'
    table_body= browser.find_element(By.XPATH, tbody)
    table_rows= table_body.find_elements(By.TAG_NAME, "tr")
```

Figure B2 (Converting data into csv file)



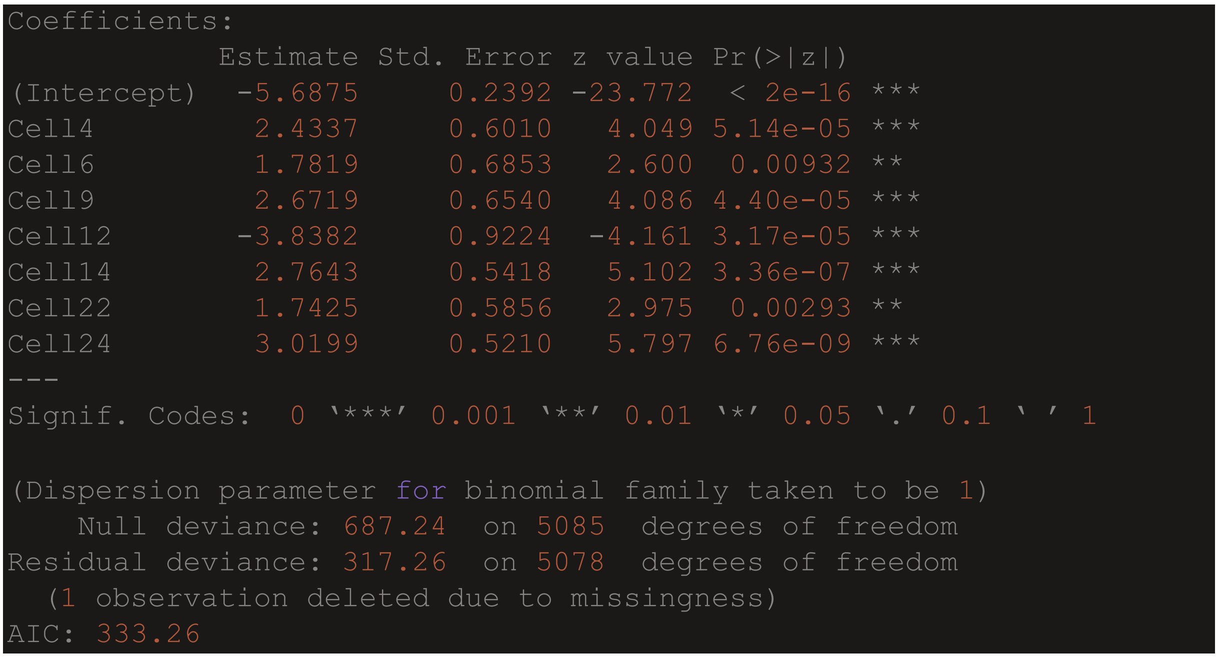
```
with open(“weatherdata.csv”,”w”) as file:
        writer = csv.writer(file, quoting=csv.QUOTE_ALL)
        for row in table_rows:
            table_data=row.find_elements(By.XPATH,’.//td[3]’)
            #table1_data=row.find_elements(By.XPATH,’.//td[1]’)
            for data in table_data:
                row_data.append(data.text)

        for j in row_data:
            writer.writerow([j])
```

Figure B3 (Logistic Regression Formulation)



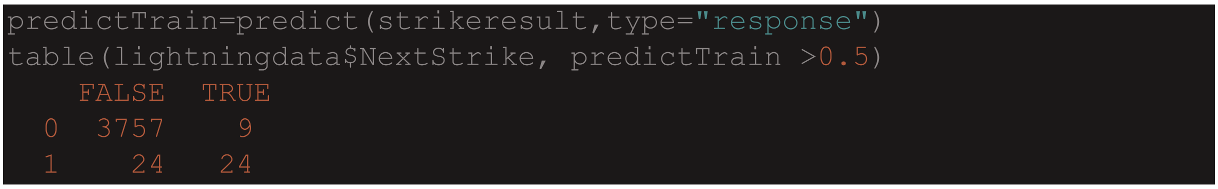
```
strikemodel<- NextStrike ~
Cell4+Cell6+Cell9+Cell12+Cell14+Cell22+Cell24
strikeresult<-glm(strikemodel, family=binomial,
data=lightningdataTrain)
```

Figure B4 (Logistic Regression Model)



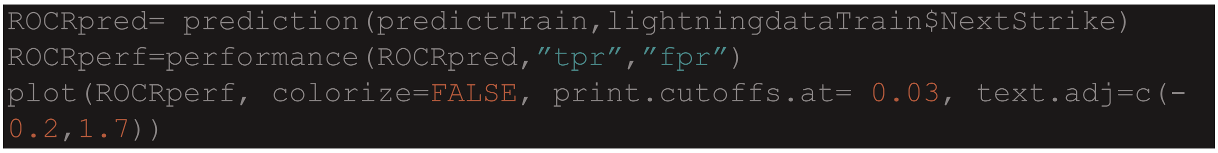
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.6875     0.2392 -23.772  < 2e-16 ***
Cell4         2.4337     0.6010   4.049 5.14e-05 ***
Cell6         1.7819     0.6853   2.600  0.00932 **
Cell9         2.6719     0.6540   4.086 4.40e-05 ***
Cell12       -3.8382     0.9224  -4.161 3.17e-05 ***
Cell14        2.7643     0.5418   5.102 3.36e-07 ***
Cell22        1.7425     0.5856   2.975  0.00293 **
Cell24        3.0199     0.5210   5.797 6.76e-09 ***
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 687.24  on 5085  degrees of freedom
Residual deviance: 317.26  on 5078  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 333.26
```

Figure B5 (Construct confusion matrix with probability threshold p = 0.5)



```
predictTrain=predict(strikeresult,type="response")
table(lightningdata$NextStrike, predictTrain >0.5)
    FALSE  TRUE
  0  3757     9
  1    24    24
```

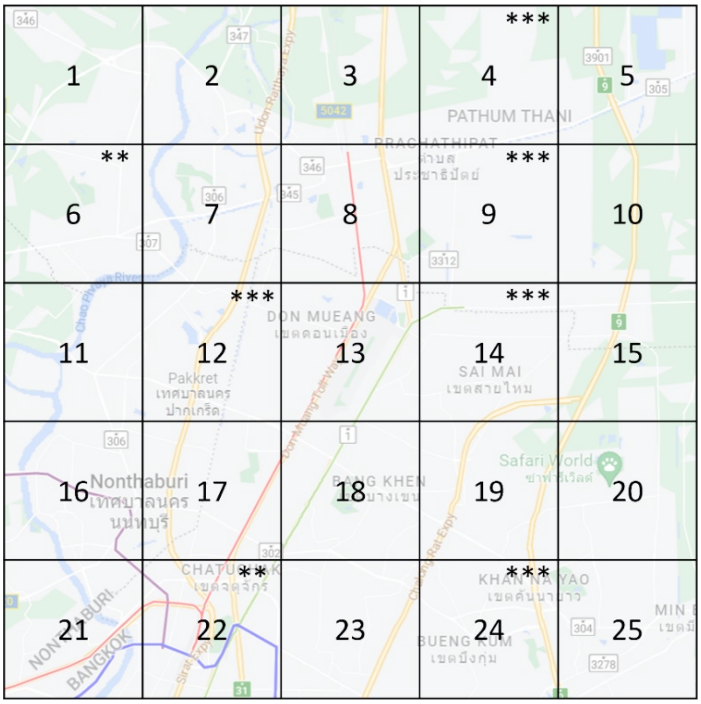
Figure B6 (Construct ROC Curve)



```
ROCRpred= prediction(predictTrain,lightningdataTrain$NextStrike)
ROCRperf=performance(ROCRpred,”tpr”,”fpr”)
plot(ROCRperf, colorize=FALSE, print.cutoffs.at= 0.03, text.adj=c(-
0.2,1.7))
```

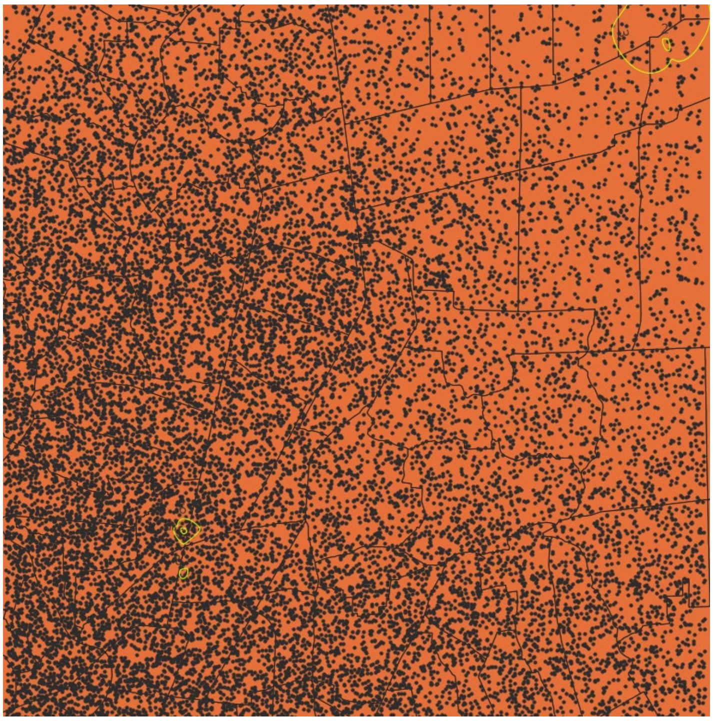
Appendix C – Tables, Graphs, Equations & Data Visualizations

Figure C1 (5x5 grid consisting of 5.5km by 5.5km squares with each square labelled from 1 to 25)



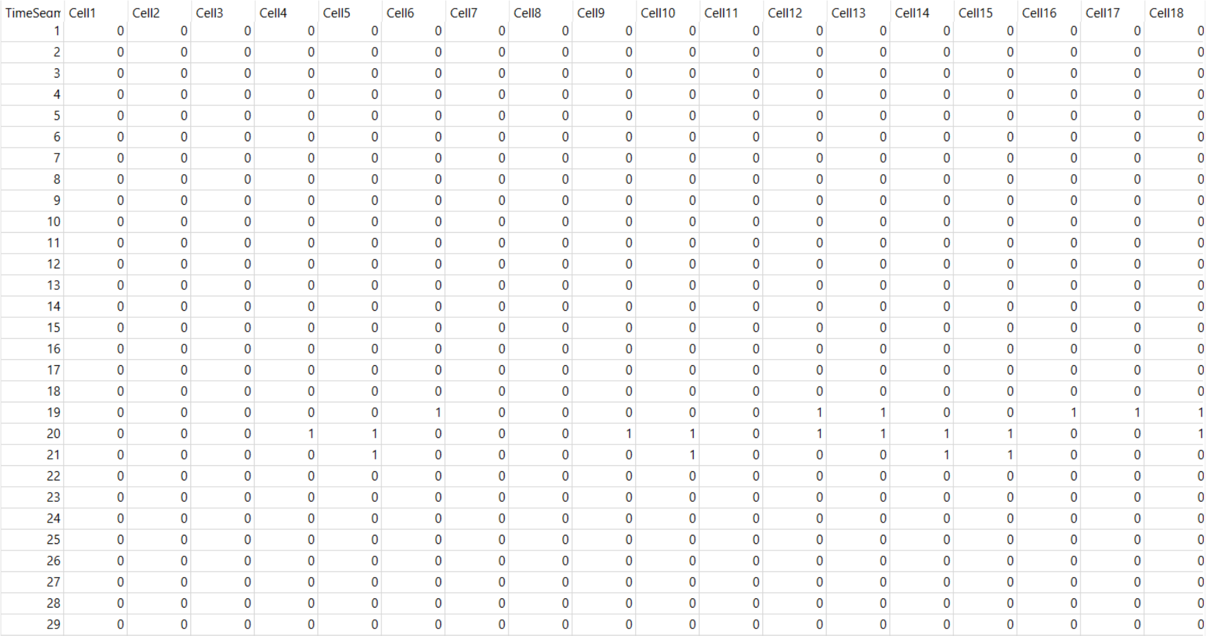
: 99% significant *: 99.9% significant

Figure C2 (Elevation & Lightning Data visualized over DMK)



Yellow markings: Contour plots around the Bangkok Metropolitan Area

Figure C3 (lightningstrikedata.csv)



TimeSean	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6	Cell7	Cell8	Cell9	Cell10	Cell11	Cell12	Cell13	Cell14	Cell15	Cell16	Cell17	Cell18
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	1	1
20	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1	0	0	1
21	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure C4 (Correlation Plot for Lightning Data)

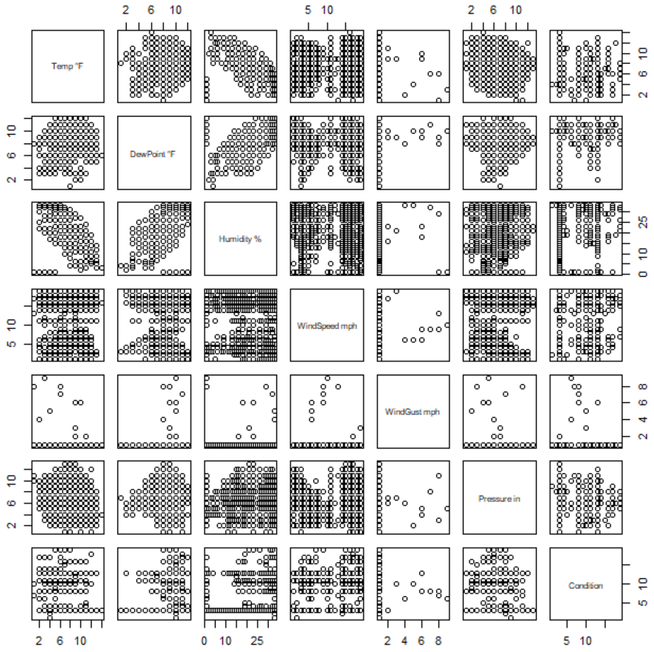
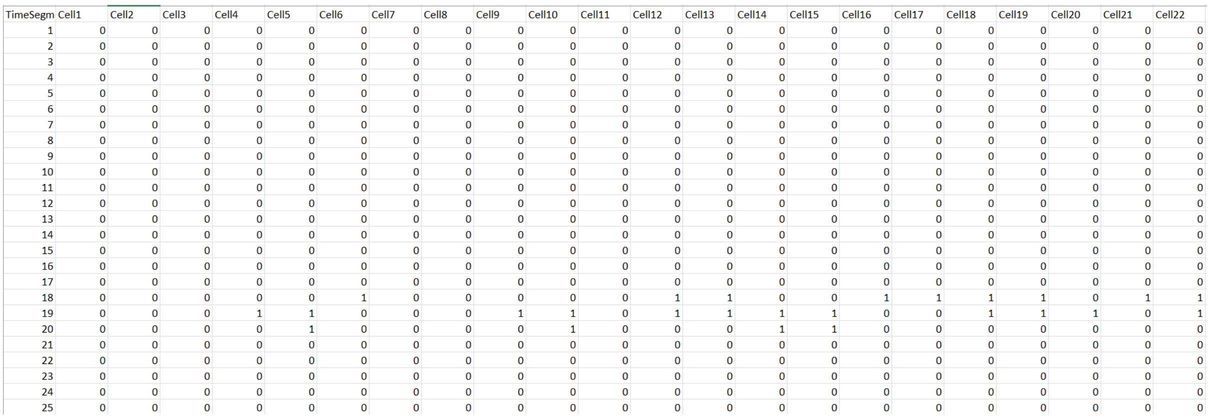


Figure C5 (Preparing the data to use in logistic regression model)



TimeSegm	Cell1	Cell2	Cell3	Cell4	Cell5	Cell6	Cell7	Cell8	Cell9	Cell10	Cell11	Cell12	Cell13	Cell14	Cell15	Cell16	Cell17	Cell18	Cell19	Cell20	Cell21	Cell22
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	1	1	1	0	1	1
19	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1	0	0	1	1	1	0	1
20	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure C6 (ROC Curve with optimal probability threshold)

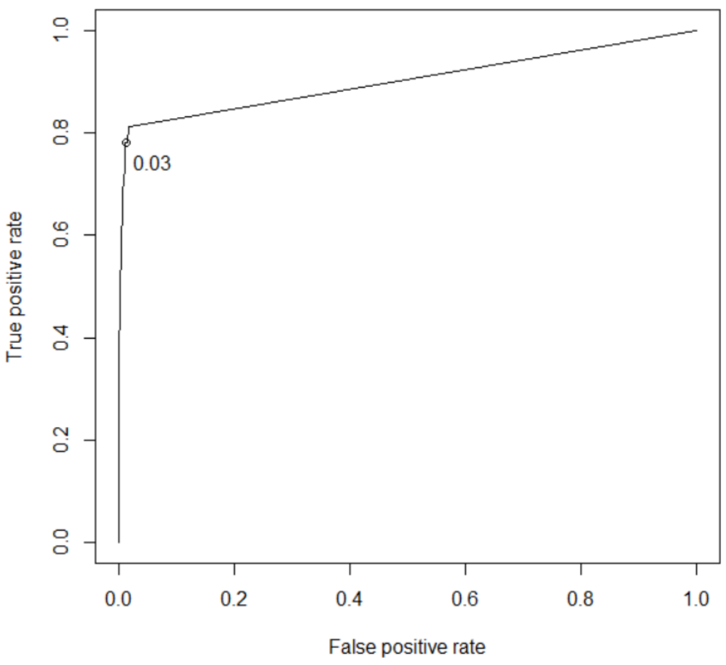
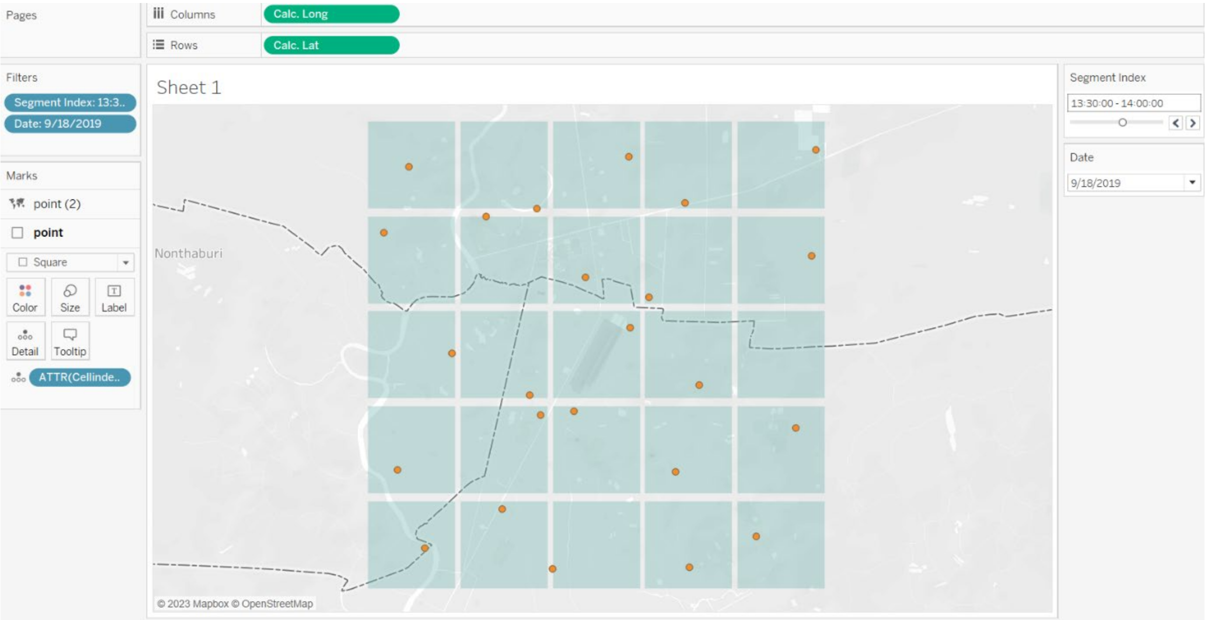


Figure C7 (Grid visualization on Tableau)



Pages

Columns: Calc. Long

Rows: Calc. Lat

Filters

Segment Index: 13:3..

Date: 9/18/2019

Marks

point (2)

point

□ Square

Color Size Label

Detail Tooltip

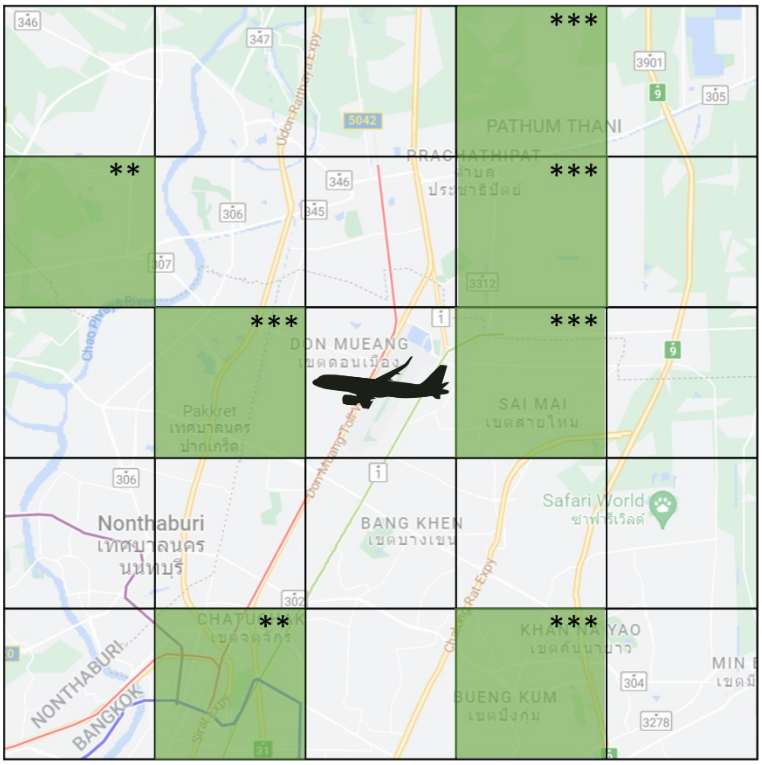
ATTR(Cellinde..

Sheet 1

Segment Index: 13:30:00 - 14:00:00

Date: 9/18/2019

Figure C8 (Grid with the selected cells in the model and its significance)



: 99% significant   *: 99.9% significant

Figure C9 (Formulations for 1h model)

Formulation No.	Variables Considered in Equation	AIC Value	All Variables Significance ≥ 95%
1	NextStrike1h ~ Cell1+Cell2+Cell3+Cell4+Cell5+Cell6+Cell7+Cell8+Cell9+Cell10+Cell11+Cell12+Cell13+Cell14 +Cell15+Cell16+Cell17+Cell18+Cell19+Cell20+Cell21+Cell22+Cell23+Cell24+Cell25+Temp+ DewPoint+Humidity+WindGust+WindSpeed+Pressure	314.27	No
2	NextStrike1h ~ Cell6+Cell8+Cell12+Cell14+Cell24+Cell25	319.32	No
3	NextStrike1h ~ Cell8+Cell12+Cell14+Cell24+Cell25	320.74	Yes

Figure C10 (Formulations for 30min model)

Formulation No.	Variables Considered in Equation	AIC Value	All Variables Significance ≥ 95%
1	NextStrike30m ~ Cell1+Cell2+Cell3+Cell4+Cell5+Cell6+Cell7+Cell8+Cell9+Cell10+Cell11+Cell12+Cell13+Cell14 +Cell15+Cell16+Cell17+Cell18+Cell19+Cell20+Cell21+Cell22+Cell23+Cell24+Cell25+Temp+ DewPoint+Humidity+WindGust+WindSpeed+Pressure	339.68	No
2	NextStrike30m ~ Cell2+Cell4+Cell6+Cell9+Cell12+Cell14+Cell16+Cell19+Cell22+Cell23+Cell24	332.62	No
3	NextStrike30m ~ Cell4+Cell6+Cell9+Cell12+Cell14+Cell22+Cell24	333.26	Yes

Figure C11 (Logistic Regression Equation of model for 1hr and 30 mins)

$$\boldsymbol{P(Lightning\ Strike\ at\ Airport\ [1h])} = \frac{1}{1+e^{-4.7735+2.2982*Cell8+1.4949*Cell9+1.7013*Cell14-2.1315*Cell15+2.6276*Cell24}}$$

$$\boldsymbol{P(Lightning\ Strike\ at\ Airport\ [30m])} = \frac{1}{1+e^{-5.6875+2.4337*Cell4+1.7819*Cell6+2.6719*Cell9-3.8382*Cell12+2.7643*Cell14+1.7425*Cell22+3.0199*Cell24}}$$

Figure C12 (Confusion Matrices for 1hr and 30 min models)

Don Mueang International Airport (1 Hour)		Predicted	
		No Lightning	Lightning
Actual	No Lightning	**2445**	49
	Lightning	18	**31**

Don Mueang International Airport (30 mins)		Predicted	
		No Lightning	Lightning
Actual	No Lightning	**4964**	58
	Lightning	14	**50**

Figure C13 (Comparison of Accuracy, Specificity & Sensitivity for 1hr and 30 mins)

	Don Mueang International Airport (1 Hour)	Don Mueang International Airport (30 Mins)
Accuracy	97.71%	98.58%
Specificity	98.04%	98.84%
Sensitivity	63.27%	78.13%

Figure C14 (Confusion Matrices of JIA & KLIA)

Juanda International Airport (30 mins)		Predicted	
		No Lightning	Lightning
Actual	No Lightning	**5041**	21
	Lightning	11	**14**

Kuala Lumpur International Airport (30 mins)		Predicted	
		No Lightning	Lightning
Actual	No Lightning	**4892**	98
	Lightning	37	**60**

Figure C15 (Comparison of Accuracy, Specificity & Sensitivity of DMK, JIA, KLIA)

	Don Mueang International Airport (30 Mins)	Juanda International Airport (30 Mins)	Kuala Lumpur International Airport (30 Mins)
Accuracy	98.58%	99.37%	97.35%
Specificity	98.84%	99.56%	98.04%
Sensitivity	78.13%	56.00%	61.86%

Figure CX (Confusion Matrix)

Confusion Matrix		Predicted	
		No Lightning	Lightning
Actual	No Lightning	**TN**	FP
	Lightning	FN	**TP**