Predicting Rainfall Amount in the State of Utah for 3 Different Time Scopes Using Supervised Learning Classification Models

Lecio Ribeiro

CS180

Doctor Wingate

December 15th, 2023

Contents

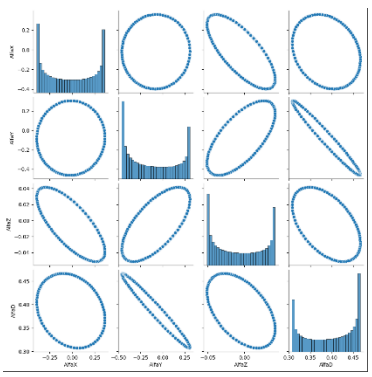
Predicting Rainfall Amount in the State of Utah for 3 Different Time Scopes Using Supervised Learning Classification Models

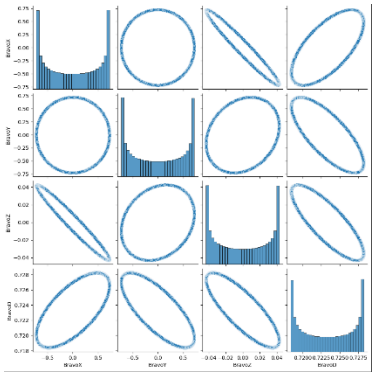
- ## Exploratory Data Analysis

The goal of this research is to answer the following question: "Is it possible to predict rainfall amount in the state of Utah using the Input-Targets dataset?" However, before doing inference procedures with that data, it is important to process it and analyze it with typical statistical methods that give relevant insights into the most efficient paths to follow when making classification predictions. Other than that, calculating the empirical probability of one of the 5 possible rain categories that classify instances in the following time periods: 7 days, 14 days, and 28 days, will measure the uncertainty of those variables which can be compared with the Machine Learning models predictions in the results section.

Analyzing Relationships and Forms of Variables

Without the technical knowledge to understand the features from the dataset and limited computer memory for checking the relationship of different features, the pairplot method from seaborn becomes useful for visualizing the form of variables relationship at once. So, by randomly picking 10 feature variables to analyze with pairplot 4 times, it was concluded (at least based on the limited computations and amount of this from this project) that the only features that were somehow correlated were the ones with the same initial name (eg., "Alpha", "Bravo", etc.). Their forms were all circular, varying from a perfect circle to oval shapes. To understand their relationship better, the pairplot method was then used in each of those feature groups separately (see attached image below of Alpha and Bravo).





Considering the strange/undetermined behavior from circular relationships, the complexity of matching machine learning classification models to this type of data, and the lack of relationship between features out of their phonetic classification[1], the next research steps will train ML[2] models using the features:

1. Within the same phonetic classification
2. From the relationships that have an oval shape instead of a circular shape
3. Following the previous requirements and that are also more efficient in predicting the rain type classifications based on what is outlined on the Test/Train Split section.

[1]The Phonetic Alphabetic is constituted by Alpha, Bravo, and the other feature classifications from the "Input-Targets" dataset

[2] Abbreviation for Machine Learning

Calculating Empirical Probabilities of Each Rain Category for Each Time Scope

The empirical probabilities of each category on each time scope were calculated by diving the number of instances classified as certain rain category[3] by the number of instances from the dataset. The results are the following:

```
for i in ['7P', '14P', '28P']:
  for j in range(1,6):
    new_df = df[df[i] == j]
    print("TIME SCOPE = ", i, "\n", "RAIN TYPE = ", j, "\n", "prob. = ", len(new_df)/len(df))
```

```
TIME SCOPE =  7P
 RAIN TYPE =  1
 prob. =  0.276199031974186
TIME SCOPE =  7P
 RAIN TYPE =  2
 prob. =  0.19587855676151364
TIME SCOPE =  7P
 RAIN TYPE =  3
 prob. =  0.1452955412144324
TIME SCOPE =  7P
 RAIN TYPE =  4
 prob. =  0.16155764153710767
TIME SCOPE =  7P
 RAIN TYPE =  5
 prob. =  0.22106922851276034
TIME SCOPE =  14P
 RAIN TYPE =  1
 prob. =  0.10351276034027575
TIME SCOPE =  14P
 RAIN TYPE =  2
 prob. =  0.2668304488119683
TIME SCOPE =  14P
 RAIN TYPE =  3
 prob. =  0.17532634203578762
TIME SCOPE =  14P
 RAIN TYPE =  4
 prob. =  0.24453652097389264
TIME SCOPE =  14P
 RAIN TYPE =  5
 prob. =  0.209793927838075 68
TIME SCOPE =  28P
 RAIN TYPE =  1
 prob. =  0.019415517747139923
TIME SCOPE =  28P
 RAIN TYPE =  2
 prob. =  0.23573628630096802
TIME SCOPE =  28P
 RAIN TYPE =  3
 prob. =  0.23102449398650632
TIME SCOPE =  28P
 RAIN TYPE =  4
 prob. =  0.19945365209738927
TIME SCOPE =  28P
 RAIN TYPE =  5
 prob. =  0.3143700498679965
```

Based on the results, there is no clear trend between the probabilities, but 28P, rain type 1, and 14P, rain type 1 have the smallest probabilities, indicating that rain type 1 is rare within these time periods.

- # Test/Train Split

This part of the research process determines the feature space used for training the ML models based on the requirements underlined on the subsection: ”Analyzing Relationships and Forms of Variables” along with new considerations determined on the next section. Other than that, it describes the split of the dataset between what will be used to train the machine learning models and what will be used to test their efficiency later along with justifying the split methods.

Feature Space

Other than following the 3 requirements from the “Analyzing Relationships and Forms of Variables” subsection, the feature space will be determined by getting the features from the phonetic classification that in average, are the closest to a linear form. To measure that, the average r^2 was taken from all the possible plots for each phonetic category. Interestingly, all the phonetic categories presented the same average r^2, which was approximately 0.32937924353593445. Because of that, choosing the features with oval shape (biggest r^2) within any of the phonetic categories should all be as efficient as each other. This led the feature space to be AlfaY and AlfaD, since (within the Alfa phonetic category) their relationship is the one with the biggest r^2.

[3] The rain categories were labeled as 1 to 5

Split Proportions

To assure that the models do not overfit and do not underfit, the chosen split proportion was the standard one for most ML models training: 80% of the data for training and 20% of the data for testing the efficiency of the models.

- ## Training the ML models

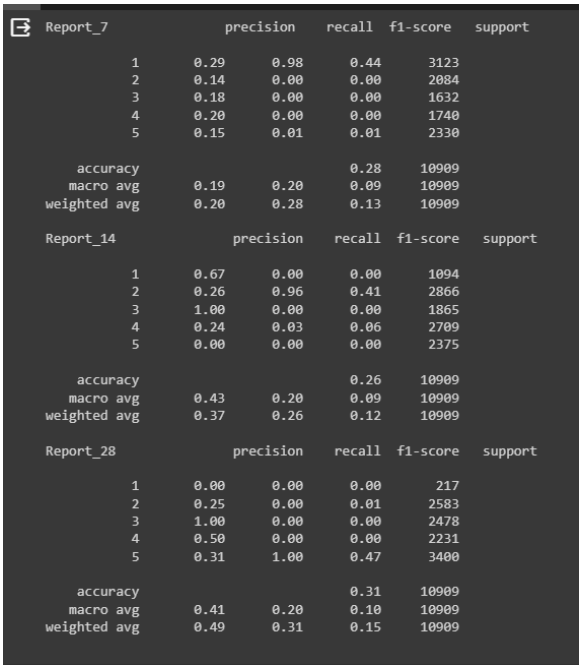
This step is the last layer of work before getting the research results and consists of following certain guidelines to decide the most efficient ML model to use in this research process and analyzing the performance of each model analyzed.

Model Selection and Efficiencies

Considering the tradeoff for precision and recall, the priority for the scenario of this research is getting the highest accuracy and recall instead of precision since the priority is to avoid misclassifying the rain type for the different time scopes. Therefore, the model with the highest average recall for each rain type category plus accuracy will be the chosen one to be used for predicting rainfall amount in the state of Utah. Also, it is important to point out that ROC curves are not being used in this case because not all models have stochastic outputs. Other than that, to assure that the models are giving their best performance when compared to the others, the models will also be trained with a variety of parameters and hyperparameters and only the highest accuracy + average recall will be considered for each model.

AdaBoostClassifier

This model performed extremely poorly by classifying all instances as one of the 5 categories. Also, this behavior did not change independently of the parameters and hyperparameters.



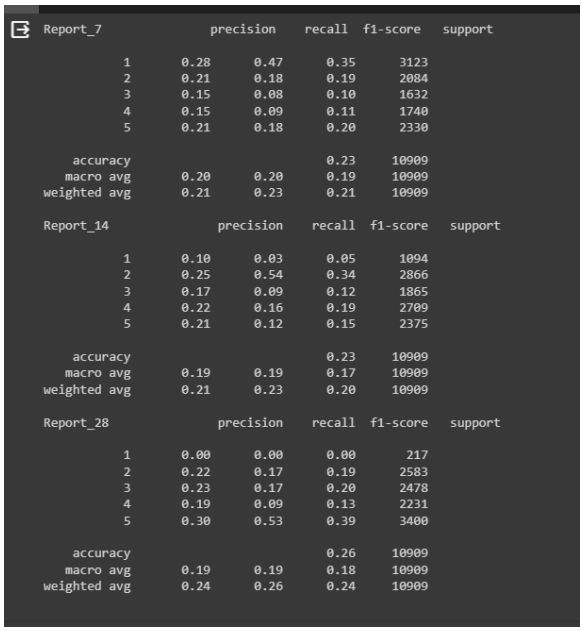
Report_7	precision	recall	f1-score	support
1	0.29	0.98	0.44	3123
2	0.14	0.00	0.00	2084
3	0.18	0.00	0.00	1632
4	0.20	0.00	0.00	1740
5	0.15	0.01	0.01	2330
accuracy			0.28	10909
macro avg	0.19	0.20	0.09	10909
weighted avg	0.20	0.28	0.13	10909

Report_14	precision	recall	f1-score	support
1	0.67	0.00	0.00	1094
2	0.26	0.96	0.41	2866
3	1.00	0.00	0.00	1865
4	0.24	0.03	0.06	2709
5	0.00	0.00	0.00	2375
accuracy			0.26	10909
macro avg	0.43	0.20	0.09	10909
weighted avg	0.37	0.26	0.12	10909

Report_28	precision	recall	f1-score	support
1	0.00	0.00	0.00	217
2	0.25	0.00	0.01	2583
3	1.00	0.00	0.00	2478
4	0.50	0.00	0.00	2231
5	0.31	1.00	0.47	3400
accuracy			0.31	10909
macro avg	0.41	0.20	0.10	10909
weighted avg	0.49	0.31	0.15	10909

DecisionTreeClassifier

This model performed with similar accuracy to the last model but it outputted different categories for the instances. Also, the parameters that led the model to perform its best were the following: DecisionTreeClassifier(random_state=0, max_depth= 60, min_samples_split = 23).



Report_7	precision	recall	f1-score	support
1	0.28	0.47	0.35	3123
2	0.21	0.18	0.19	2084
3	0.15	0.08	0.10	1632
4	0.15	0.09	0.11	1740
5	0.21	0.18	0.20	2330
accuracy			0.23	10909
macro avg	0.20	0.20	0.19	10909
weighted avg	0.21	0.23	0.21	10909

Report_14	precision	recall	f1-score	support
1	0.10	0.03	0.05	1094
2	0.25	0.54	0.34	2866
3	0.17	0.09	0.12	1865
4	0.22	0.16	0.19	2709
5	0.21	0.12	0.15	2375
accuracy			0.23	10909
macro avg	0.19	0.19	0.17	10909
weighted avg	0.21	0.23	0.20	10909

Report_28	precision	recall	f1-score	support
1	0.00	0.00	0.00	217
2	0.22	0.17	0.19	2583
3	0.23	0.17	0.20	2478
4	0.19	0.09	0.13	2231
5	0.30	0.53	0.39	3400
accuracy			0.26	10909
macro avg	0.19	0.19	0.18	10909
weighted avg	0.24	0.26	0.24	10909

GaussianProcessClassifier

This model presented two major challenges:

1. It would consume all the RAM if trained with all the instances from the feature space, so it was necessary to sample the feature space with 1000 instances.
2. Time for training. Since training would take in average 5 minutes each, not many parameters and hyperparameters were compared

Other than that, this model also did not perform as expected, and mostly categorized most instances as just one category.

Multi-Layer Neural Net

Unfortunately, this model also did not perform well under any parameter and its behavior was like the previous models except for decision trees. This may be the key indicator that the previous methodologies used for preparing the models training did not work as expected, which may mean that the whole research process failed to either prove or disprove the main hypothesis. This will be better disserted in the "Conclusion" section.

• Results

**Not relevant based on training results unfortunately*

7P

14P

28P

• Conclusion

Before giving conclusions about the research results it is important to consider that the research could have achieve better and more efficient outcomes if the following limitations did not exist:

1. Project Time Constrain (more models, parameters, hyperparameters, splits, features could have been tested and the dataset would have been more wrangled)
2. Computer RAM and Graphics card memory limitation (with more memory it would be possible to train the ML models with more data)

Other than that, the methodology followed by this research first constrained guidelines for the feature space and calculated the empirical probability of each type of rain for each time scope analyzed. Then, it defined more guidelines for the feature space and splitting the data used to train and test the models. Finally, it trained 4 ML models and defined the most efficient of them prioritizing the highest average recall for the possible categories.

Is it Possible to Predict Rainfall Amount in the State of Utah Using the Dataset?

In conclusion, since the research methodology failed to prepare the data enough so machine learning models could predict the rain categories, the main contribution of the research is its explanatory data analysis. This part of the research was still methodical enough to generate relevant results that contained relevant insights about the nature of the "Inputs'Target" dataset and its capability of being used to predict rainfall amount in the state of Utah. In other words, the machine learning models may have failed to predict rainfall amount in the state of Utah by following the research methodology, but the research still gives evidence that this is still possible by following other methodologies and other statistical techniques.