Machine Learning Engineer Nanodegree

Capstone Proposal

Shankar Cherla

May 21st, 2019

Predicting if it rains tomorrow in Albury, Australia.

Domain Background

Weather Forecasting is a branch of earth science. Traditionally, it is done through physical simulations of the atmosphere. For example, weather balloons are used for weather forecasting. Its current state is sampled, and the future state is calculated using fluid dynamics and thermodynamics equations. Weather forecasting through physical simulations demands understanding of complex physical processes. However, this physical model is subject to disturbances and uncertainties in the measurements.

Thus, machine learning represents a viable alternative in the weather forecast. It is comparatively robust to disturbances, does not require a complete understanding of physical processes and the predictions are trustworthy as the statistical model developed is completely data driven.

Source:

- The weather forecast dataset is taken from Kaggle.
- The kernel which I used for reference is also from Kaggle.
- This binary classification paper is used for reference

Problem Statement

In this project I wanted to predict if it would rain tomorrow in Albury, Australia. I utilized supervised learning as I want my model to predict by learning from data. Since our target variable is a binary class variable with outcome either Yes or No, I chose to apply classification algorithms. Later on, the model can be used to predict if it rains or not for new/unseen data.

Datasets and Inputs

This dataset contains daily weather observations from numerous Australian weather stations among which I considered Albury weather station and the initial shape of the dataset is (3011, 24). After removing all the null values both column wise and row wise the final shape of dataset changes and is (2440,17). Our data has 1913 No and 527 Yes class labels for the predictor variable 'RainTomorrow'. This shows our data is skewed. The daily observations are available at http://www.bom.gov.au/climate/data Copyright
Commonwealth of Australia 2010, Bureau of Meteorology. The weather forecast dataset is taken from Kaggle.

Input information:

1. 'Date': The date of observation
2. 'Location': The common name of the location of the weather station
3. 'MinTemp': The minimum temperature in degrees celsius
4. 'MaxTemp': The maximum temperature in degrees celsius
5. 'Rainfall': The amount of rainfall recorded for the day in mm
6. 'Evaporation':The so-called Class A pan evaporation (mm) in the 24 hours to 9am
7. 'Sunshine': The number of hours of bright sunshine in the day.
8. 'WindGustDir': The direction of the strongest wind gust in the 24 hours to midnight.
9. 'WindGustSpeed': The speed (km/h) of the strongest wind gust in the 24 hours to midnight.
10. 'WindDir9am': Direction of the wind at 9am.
11. 'WindDir3pm': Direction of the wind at 3pm.
12. 'WindSpeed9am': Wind speed (km/hr) averaged over 10 minutes prior to 9am WindSpeed3pm.
13. 'Wind speed (km/hr) averaged over 10 minutes prior to 3pm
14. 'Humidity9am: Humidity (percent) at 9am

15. ‘Humidity3pm’: Humidity (percent) at 3pm
16. ‘Pressure9am’: Atmospheric pressure (hpa) reduced to mean sea level at 9am
17. ‘Pressure3pm’: Atmospheric pressure (hpa) reduced to mean sea level at 3pm
18. ‘Cloud9am’: Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
19. ‘Cloud3pm’: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm.
20. ‘Temp9am’: Temperature (degrees C) at 9am. 21. ‘Temp3pm’: Temperature (degrees C) at 3pm.
22. ‘RainToday’: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.
23. ‘RISK_MM’: The amount of next day rain in mm. Used to create response variable.
24. ‘RainTomorrow’: The target variable. Will it rain tomorrow?

Solution Statement

My intention of this project is to predict if it rains tomorrow in Albury, Australia utilizing ML. To serve the same purpose I will utilize supervised learning for developing a model. As our target variable ‘RainTomorrow’ is simply a binary class i.e., it contains either Yes or no, I opt to apply classification algorithms. I will create models by applying various classification algorithms and choose the model with greatest performance. I will try to improve the performance of the chosen model by tuning parameters, try to fit it on reduced data and calculate respective scores in each scenario. In this way after all the steps I will choose the best model suitable with best performance and later this model can be used to predict if it rains or not for new/unseen data.

Benchmark

For the benchmark model, I used Gaussian Naïve Bayes algorithm from scikit learn module. I used Accuracy and F-score as evaluation metrics. As the data is skewed, I preferred to use F-score for evaluation of performance. As F-score is performance evaluation metric, this score of Benchmark model will be compared with the respective scores of the new models. The purpose of this comparison of scores is to check how better the new models are good at prediction when compared with benchmark model.

Evaluation Metrics

For Classification problems there are many types of performance evaluation metrics namely Accuracy, F-score etc. Our data has 1913 No and 527 Yes class labels for the predictor variable 'RainTomorrow'. This shows our data is skewed. For classification problems that are skewed in their classification distributions like in our case, Accuracy is not dependable measure for performance evaluation and hence precision, recall are very useful. These two metrics can be combined to get the F-beta score, which is weighted average of the precision and recall scores. This score can range from 0 to 1, with 1 being the best possible score. In particular, when beta = 0.5, more emphasis is placed on precision. This is called the F-0.5 score or F-score for simplicity.

The general formula of F-score for positive real β is:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

Source:

There are many articles giving detailed explanation for choosing the right performance evaluation metric. I am linking the article which I referred to.

Project Design

The theoretical workflow I planned to approach to the solution for Capstone project is given below.

Data Acquisition, Exploring, Plotting and Preprocessing:

- Firstly, I will acquire dataset from Kaggle.
- I will explore to get to know about the columns of dataset and the type of data we are dealing with.
- I will remove the unnecessary columns as they constitute very less percent of whole data.
  These columns are

  'Sunshine','Evaporation','Cloud3pm','Cloud9am','Location','RISK_MM','Date'

- Null values if any are removed from the dataset.
- I will use basic plots to get to know a bit more about the data we are dealing with i.e., for analyzing our data.
- The outliers are removed using Z-score from stats of scipy module. This reduces the shape of data from (2440, 17) to (2333, 17).

- The categorical columns 'WindGustDir', 'WindDir3pm', 'WindDir9am' are converted to numerical using one-hot encoding. As the columns 'RainToday', 'RainTomorrow' contain only Yes/No they are replaced with 1/0 using lambda function. This step is implemented because the algorithms I plan to apply to develop model might not be able to deal with categorical columns.
- A Min-Max scaler is applied to data to prevent the influence of one attribute on the prediction outcome and to restrict values between certain range.

- The benchmark-model is trained on this data and the evaluation metric chosen is applied to benchmark model (for comparing benchmark model's performance score with other models). In our project F-score is preferred to Accuracy as our data is skewed.

Deciding the model:

- I plan to apply 3 classification algorithms.
- The F-score, Accuracy is computed for all the models. The best model is chosen based on its respective F-score as we are dealing with skewed data.
- In our project I will apply various classification algorithms which therefore creates various models that fit our data.
- We then generate simple plots for each respective model to aid analysis and for choosing the best model based on F-score criteria.

- The output of this step is an unoptimized model which has highest F-score when compared with other classification models.

Parameter Tuning using GridSearchCV:

To tune parameters and improve the performance of unoptimized model I will use GridSearchCV from sklearn.model_selection

- The model chosen i.e., the unoptimized model will be tuned with its respective parameter combinations resulting in an optimized model. The performance of optimized model is greater than unoptimized model.

- The main purpose of this step is to improvise our chosen model i.e., the unoptimized model a bit more using the best possible combination values of its respective parameters. This step creates an optimized model.

Knowing Feature importances :

- We use AdaBoostClassifier from sklearn.ensemble as it has feauture_importances attribute in this step.
- We plot the five most important features of our data
- We apply our optimized model on this reduced data and find respective F-score •
  We then decide whether we should use the reduced data or full data.

At end I will plot a graph depicting how accuracy and F-score varied for the models in our project i.e., benchmark model, unoptimized model (the model with best performance among all classification algorithms), optimized model (attained after tuning unoptimized model) and optimized model on reduced data.