Tyler Anderson

DSC630-T301

1/14/19

**2.2 Course Project: Milestone 1: Team Information and Project Proposal**

**Introduction**

For this project id like to explore weather data from numerous Australian weather stations. This dataset contains 10 years of daily weather observations from 2008-2017 from the Australian weather stations. This dataset is very complete and is meant for analysis. Possible cleaning has already been done prior to the release of this dataset. While there are NA values, they are selective and are meant to imply that not all datapoints hare intended to have complete data.

**Problem Statement**

My goal is to determine the likelihood of it raining the next day based on the data provided. The data is relatively easy to understand with good labels and a large dataset. I expect to be able to answer user-driven questions based on a regression model that predicts the weather for the next day.

**User-driven:**

What is the likelihood of it raining the next day? What plays a large role in the determining factor in predicting rain?

If it rains in the current day how does that affect the likelihood it will rain tomorrow?

Does the amount of precipitation indicate rain the following day?

How much precipitation is needed for an accurate measurement of rain the day of and the next day?

**Scope**

The scope of this data is large and important. All the variables in this dataset can be used to help determine the portability of rain the next day. It’s important to focus on the main question of rain probability the next day and not to explore too far into other avenues like probability of sunshine or cloudy days. The challenge is to gain an accuracy that’s good enough to report and to deliver in a way that’s understandable to meteorologists and regular users.

**Document Overview**

The data is a comma separated value (CSV) file that includes weather data from 2008-2017 across multiple cities and has 142,194 lines of data.

The sheet includes 24 variables:

Date: The date of observation

Location: The common name of the location of the weather station

MinTemp: The minimum temperature in degrees Celsius

MaxTemp: The maximum temperature in degrees Celsius

Rainfall: The amount of rainfall recorded for the day in mm

Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am

Sunshine: The number of hours of bright sunshine in the day.

WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight

WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight

WindDir9am: Direction of the wind at 9am

WindDir3pm: Direction of the wind at 3pm

WindSpeed9am: Wind speed (km/hr) averaged over 10 minutes prior to 9am

WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 3pm

Humidity9am: Humidity (percent) at 9am

Humidity3pm: Humidity (percent) at 3pm

Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am

Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm

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.

Cloud3pm: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values

Temp9am: Temperature (degrees C) at 9am

Temp3pm: Temperature (degrees C) at 3pm

RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

RISK\_MM: The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

RainTomorrow: The target variable. Did it rain tomorrow?

**Preliminary Requirement**

The dataset does not require too much preparation before analyzing allowing me to work on the analysis immediately. I am fully confident that the data provides enough information to answer my questions regarding rain the next day. The only issue is the accuracy of the data. The data comes from weather stations who might not provide fully accurate measurements. I can only trust the data is accurate and allows me to confidently answer my questions.

**Technical Approach: Analysis**

To start I will need to asses the weather data and learn more in depth about how the data is structured. I will use a SWOT analysis and focus on rain days to determine the likelihood of rain the next day.

The first step of my analysis is to explore the data and do data visualization to take a better look at the structure of the data. Looking at the shape, head and column names to start will allow me to know what I’m working with. Looking at the results and studying the data we need to drop the Risk\_MM variable, this variable may interfere with my logistic regression skewing my results and basically “cheating” the prediction by telling the model the answer before it runs. If we want to use this model for future predictions this variable must be taken out.

Looking at the data we can see there are some missing values that I will have to remove later before I’m done with my analysis. Next, I perform a Bivariate Analysis, I segregate the dataset into two variables categorical and numerical since there are both in the dataset. Lastly, I want to view the statistical properties of the dataset using describe. This helps us view the statistical properties of numerical variables and It excludes character variables.

**Technical Approach: Requirement Development**

I will use Python and R but lean more towards python in this instance and use R for exploratory analysis and visualizations. Ill use PowerPoint for my final presentation but if in feeling fancy ill use Prezi for a presentation.

**Technical Approach: Model Deployment**

The goal is to use test and train sets on our logistic regression model. Before I can suggest using such a model for real world predictions, I need to make sure it’s an accurate model and that it gets tested in our train and test sets properly.

**Technical Approach: Testing and Evaluation**

For testing I will create variables that will be easy to plug in for testing and training sets. These will be cleaned variables that are specifically used for my testing a training set for my model testing. Once this has been tested and we can achieve a high accuracy we can confidently say we can use this model in the real world.

**Expected Results**

I expect the results to be good enough to be used in the real world. I am hoping for an accuracy in the 80% range. If this is the case, we can use this model for real world weather predictions involving rain prediction.

**Management Approach: Project Plan**

For our project plan I will follow a CRISP-DM framework. I have an intermediate understanding of the data as man of the variables are easy to understand, and the goals are clear. To better understand the data, I will use data visualization and exploratory data analysis to gain better insight into what the data is telling me. I plan to employ univariate and Bivariate analysis that looks at types of variables and explore the categorical variables. I want to explore problems with the categorical variables if there are some as well as the numerical variables in the dataset. My target variable will be RainTomorrow variable and I wish to employ a logistic regression model to predict rain the next day. I plan to include a confusion matrix on top of precision, recall and f1 score. I will get inspiration for model building from Kaggle where many models are built to make predictions.

**Management Approach: Project Risk**

The risk in this project is that the model will not be good enough to justify a real-world use of the model and that the accuracy cannot measure up to real world standards. There is also the risk of the data being incomplete or not accurate enough to apply it to real world scenarios. For this project the risks are present, but I believe the data is good enough to get us a good accuracy and f1 score.

Reference

Dataset Source: Joe Young and Adam Young   
<https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>