Rainfall Prediction

Capstone Project – Machine Learning Fundamentals

Pramod Munaweera

pramod.munaweera@dialog.lk

CONTENTS

INTRODUCTION

Weather forecasting has been an important task for humans since early ages because it affects everyday activities. Accurate weather forecasting will greatly improve efficiency of human societies and reduce losses. But weather prediction is quite difficult due to uncertain nature.

Through this project I try to make predictions on rainfall in a given day for various cities in Australia based on weather data of the day.

DATA

Dataset consists of daily weather data collected in 49 cities in Australia through 2007 – 2017.

Data dictionary:

Field	Description
Date	Date data collected
Location	Name of the city
MinTemp	Minimum temperature of the day
MaxTemp	Maximum temperature of the day
Rainfall	Today's rainfall
Evaporation	Today's evaporation
Sunshine	Today's sunshine
WindGustDir	-
WindGustSpeed	-
WindDir9am	Wind direction at 9am
WindDir3pm	Wind direction at 3pm
WindSpeed9am	Wind speed at 9am
WindSpeed3pm	Wind speed at 3pm
Humidity9am	Humidity at 9am
Humidity3pm	Humidity at 3pm
Pressure9am	Atmospheric pressure at 9am
Pressure3pm	Atmospheric pressure at 3pm
Cloud9am	Cloud coverage at 9am
Cloud3pm	Cloud coverage at 3pm
Temp9am	Temperature at 9am
Temp3pm	Temperature at 3pm
RainToday	Whether it rained today or not
RISK_MM	-
RainTomorrow	Whether it rained next day or not

METHODOLOGY

Build a binary classification model to predict whether it will rain tomorrow or not using weather data of the day.

Perform EDA on dataset to identify relationships between variables, select, drop features. Identify fields that requires imputing.

Perform feature engineering to clean, impute, transform data into the format which is ready to build the Machine learning model.

Build multiple types of classification models (GaussianNB, LogisticRegression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier) and evaluate base performance. Then, do hyperparameter tuning for the model with highest base performance to improve model performance further.

RESULTS

Random forest classifier achieved the highest accuracy of 85.75%

Because the dataset is imbalanced, F1 score would be a better matric than the accuracy.

CONCLUSION

Rainfall prediction is a difficult task to achieve due to various environmental uncertainties. The data set is quite imbalanced and would need more feature engineering and oversampling to achieve more reliable results. Intend to further improve the model.