Project:

Predicting rain for tomorrow!

Presented by:

Kulothungasagaran N

27/May/2023

Background and Problem statement

Background

- Inaccurate weather forecasts can negatively impact fishing companies leading to asset loss and higher maintenance costs.
- If ships venture into storms, there will be a loss of assets and incur higher costs of maintenance.
- On the other hand, staying back due to inaccurate forecasts could result in a loss of business to the competitors.

Problem statement

- Can we build a data-driven model to better forecast whether it will rain tomorrow so that the fishing company can optimize its operations?

Analysis Methodology (1 of 2)

Data Description and EDA

Data description

- The dataset provided contains weather measurements collected from four different places in Singapore during the period 2008-2017.

Exploratory Data Analysis (EDA)

EDA was conducted to

- Identify missing values and outliers and decide how to handle them.
- Understand the unique values for each variable to determine additional data-cleaning steps.
- Analyse the distribution of the target variable (“RainTomorrow”) to check if it is balanced.
- Measure the correlations of the recorded variables with the target variable to determine if there is “predictability”.
- Identify if there is any seasonality as the data spans multiple years.

Analysis Methodology (2 of 2)

Machine Learning Modelling

The objective of the modelling task is to predict whether it will rain the next day based on the current day's measurements.

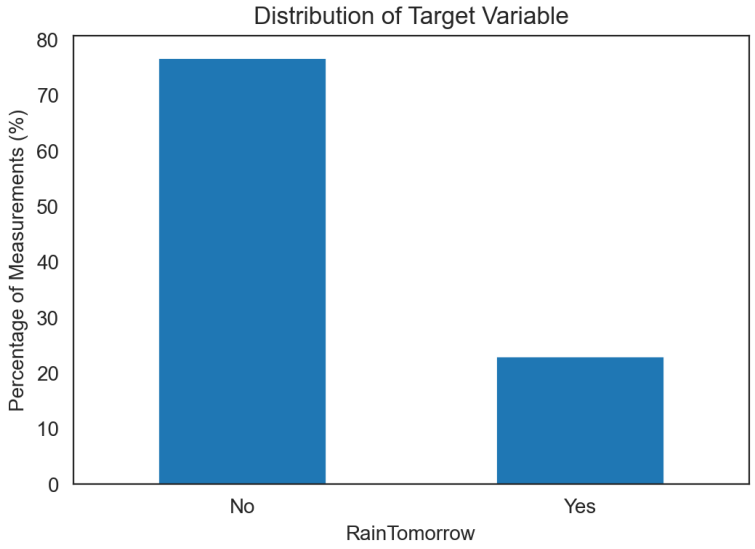
There are multiple ways to approach this task:

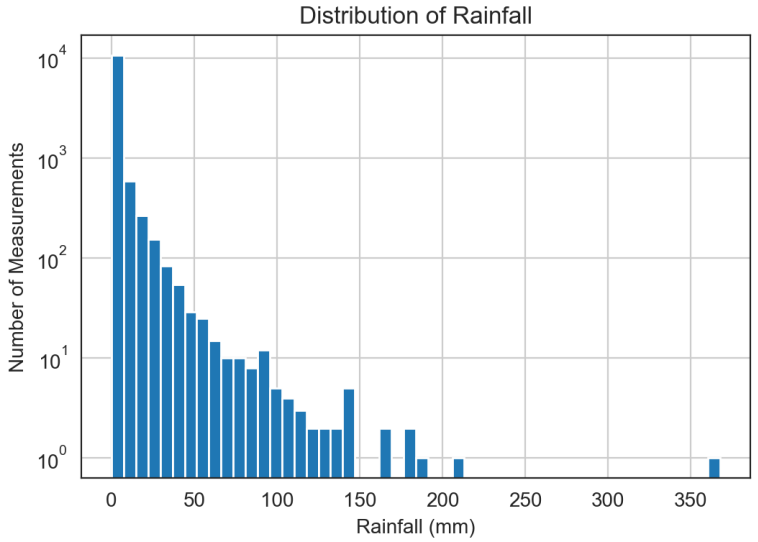
1. **Use the provided binary label directly as the target**
   - Binary classification algorithms can be directly used to predict the provided target label ("RainTomorrow") based on the other key variables recorded.
2. **Use the amount of Rainfall provided**
   - The amount of rainfall recorded for each day can be used to derive a continuous label for the rainfall measurements for the previous day.
   - Regression algorithms can be used to predict the derived rainfall amount for the next day to indicate rainfall severity.
   - Alternatively, the derived rainfall amount can be binned (Low, Med, and High rainfall) and Multiclass classification algorithms can be used to predict the rainfall severity to make fine-grained business decisions.

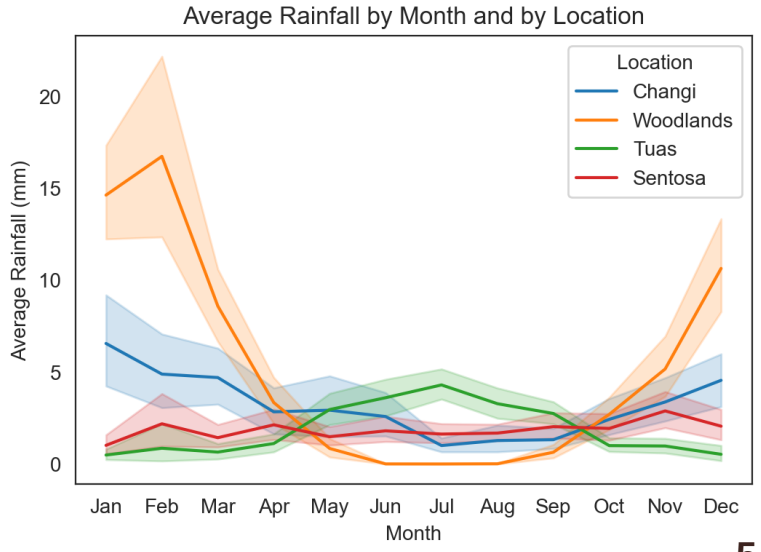
For this project, we present the results from Approach 1, as Approach 2 requires rainfall measurements every day and there were missing daily measurements for one of the locations (Sentosa).

Key EDA Insights

- Target variable ("RainTomorrow") is imbalanced → More than 70% of measurements indicate that there was no rain.
- Average rainfall shows variation with respect to Location as well as the Month of the year indicating that these two variables can be useful predictors.
- Pairplots (presented in the Appendix) reveal that the amount of humidity, sunshine, and cloud cover have a strong relationship with the target variable.



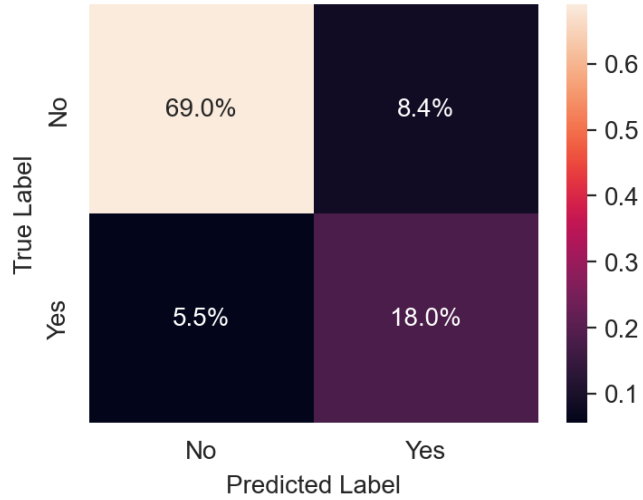




Modelling Results

- Four classification algorithms were tested. As the target variable is imbalanced, the models were evaluated based on the AUROC score.
- 5 fold cross-validation (GridSearchCV) was used to select the best model and hyperparameters. **RandomForest** came out as the best model based on the CV score.
- Decision threshold is selected after the best model is chosen based on the business criterion. From the confusion matrix we can see that ~87 % accuracy in predicting whether it will rain the next day while providing an optimum balance for the FN and FP.

No.	Algorithm	Best CV score
1	Logistic Regression	0.905
2	Random Forest	0.922
3	XGBoost	0.919
4	K Nearest Neighbors	0.896



Conclusion and next steps

- We have built an ML model with ~87 % accuracy in predicting whether it will rain the next day, while providing an optimum balance for the business.
- The fishing company can run a pilot program deploying this model for inference on next day rain predictions.

Next steps

1. Explore more on using the amount of rainfall provided to predict rainfall severity which enables the Fishing company can make fine-grained business decisions

THANK YOU