STATS 402 - Interdisciplinary Data Analysis

<Project Title>

Milestone Report: Stage 3

<Your names here>
<Your emails here>

There are no specific requirements for the stage 3 report since the progress may vary among different groups. Generally, there are four parts you need to cover in your report.

1. **Summarize your work in the past two weeks and the current status of your project**
2. **Demonstrate some results that you have obtained, and provide some essential analysis of the results.**
3. **The plan for the following final week to ensure that the project will be completed on schedule.**

Random Forest Predictions - Luka Mdivani

In the last two weeks the usage of Random Forest models for our predictive tasks has significantly expanded. But there was a slight change in one of the research questions, with the goal of avoiding repetitive code and work, instead of making two different models to predict MaxTemperature and MinTemperature. The two columns were combined to make a single AverageTemperature column, which was used as our response variable for Random Forest Regression.

Thus, our first research question was to predict average temperature using a Random Forest Regressor. Additionally, instead of using the data for each day to predict the average temperature on that same day, I decided that it was more rational and practical to try and predict the temperature on the next calendar day. I also made significant changes in the set of predictor variables. The first part of the Feature selection process was intuitive: during the data cleaning stage I got rid of all the columns who had predominantly missing/NAN rows which included :Cloud9am, Cloud3pm, Evaporation, Sunshine, RISK_MM . We also dropped Date and Location columns since they contained redundant information which could not be categorized in an efficient and useful manner. The next pre-processing stage included encoding the categorical variables in our remaining features, which were : `'WindGustDir'`, `'WindDir9am'`, `'WindDir3pm'`,`'RainToday'` and we convert the columns into One Hot representation form. For the remaining features, we further cleaned the data by filling in the NA values in each column by the mean value of the other rows in the respective columns. One additional filter which we introduced was that we dropped the `'Temp3pm'`,`'Temp9am'` columns, as our goal was to predict average temperature the next calendar day having the temperature values present introduced a bias. We used the '**SelectKBest'** method given in the Sklearn library to identify the 4 most important features for the regression model, which turned out to be : 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm'. Despite this, for the sake of gaining more information about feature importance we decided to keep existing feature variables in our model. We performed a 20%-80% test/train split on the data and trained the RandomForestRegressor() model. After the training we used the inbuilt class feature_importances of tree based classifiers, to gain more insight into what the important features were, the top 10 are visualized here in **Figure 1.**

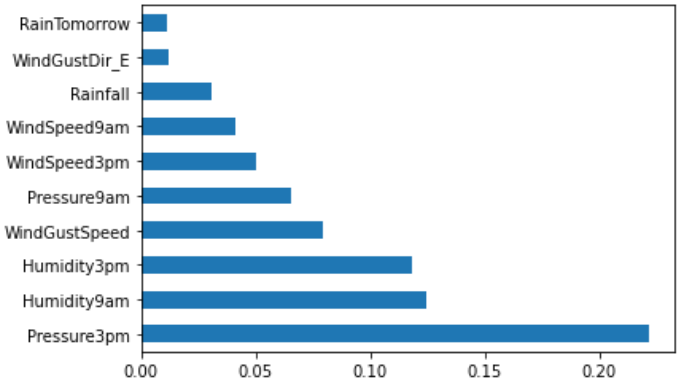

Figure 1. *Top 10 contributing features for RandomForestRegression*

We can see that the 4 features we identified using **SelectKBest** method are in the top 5 most contributing features, WinfGustSpeed being the fifth one.

We used the same data to train a second model with just the variables identified by SelectKBest as predicting features.

Models performed with varying success as the measured R2 and MSE values are displayed in the table below.

	R-Squared	MSE
Random Forest Regressor- Hand picked predictor variables	0.59	16.14
Random Forest Regressor- 4 automatically selected predictor variables	0.36	24.96

Table 1 Performance of Random Forest Regressors

As you can see from table 1, we got significantly better performance when more variables were used as predictors.

The overall relationship between predicted and true values on the test set are visualized in Figure 2 below.

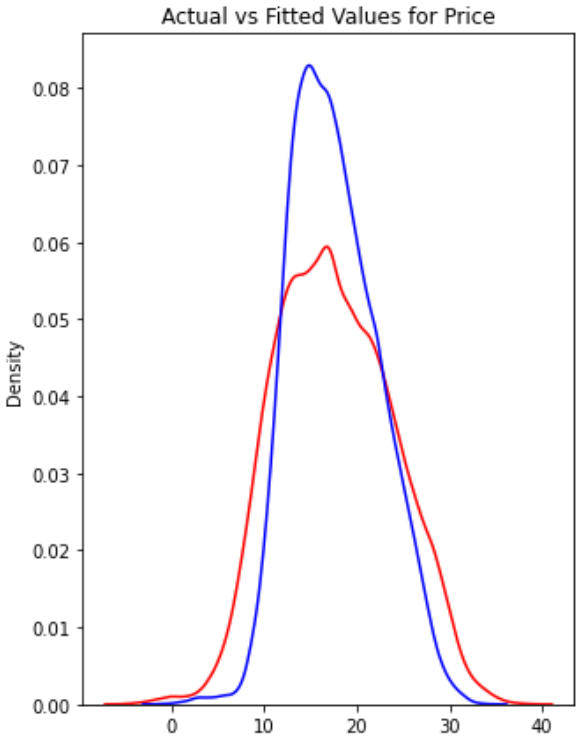

Figure 2: Predicted vs Actual test values.

The second research question was to predict whether it would rain the next day. The same data cleaning process was followed in the pre-processing stage of this model as well. Conveniently the dataset had a column of a variable called 'RainTomorrow' which stored historical data of whether it rained the next calendar day, which became our target response variable. We ran the "SelectKBest" method to find the 4 most important predictor features using the chi2 metric, which turned out to be 'Rainfall', 'Humidity3pm', 'RainToday_No', 'RainToday_Yes' . These variables were selected as our predictors

We used a standard 15%-75% train/test split, and trained the RandomForestClassifier model on the training data. The accuracy score of our model turned out to be 0.82, which is a good performance measure. The confusion matrix of our model is presented in Figure 3.

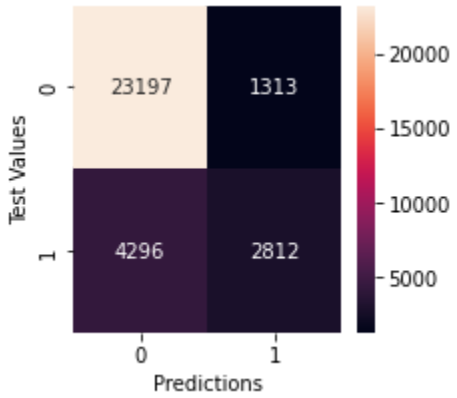

Figure 3. *Confusion Matrix.*

We also trained a second model, but this time without limiting the predictor features to those identified by the chi2 method. And we received a slightly larger accuracy of 0.84. The confusion matrix of the second model is in Figure 4.

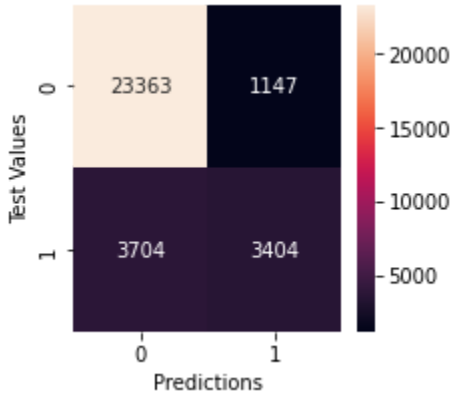

Figure 4. *Confusion Matrix Second Model.*

I will look at ways to possibly extend or tune our models, analysis and result interpretation of our model. As well as work on finalizing the report.

References

[1] Use an enumerated list here for any references, such as books or journal/conference papers.