BA 878 Machine Learning and Data Infrastructure in Healthcare
Boston University Questrom School Of Management
Professor: Ned McCague
Student/ Report By : Prateek Naharia | nahariap@bu.edu

Homework 2

Nov, 3rd, 2023

OVERVIEW

Assignment/Report is CDC mortality dataset. The Kaggle version of the dataset can be found at this link: https://www.kaggle.com/cdc/mortality. The dataset includes information about comorbidities that have contributed to each death.

Colab Notebook Link : Naharia_Prateek_code.ipynb

Problem 1. Classification

1. Data Exploration
    a. There are 198415 numbers of individuals in the dataset.
    b. 4th day of the week or Wednesday is the day when death occur for the decedent #62688 & death cause code is 38 & as per the Kaggle data dictionary it is Motor vehicle accidents.
2. Binary Classification
    a. Trained Random Forest Classifier with max_depth = 10 & random_state=7 & Logistic Regression with random_state=7 & max_iterations=1000, fitting the data with X_train & train_target.
    b. Computing Metrics on the test set, quick summary:

	Random Forest	Logistic Regression
Accuracy	0.7466	0.7455
Precision	0.7127	0.7316
Recall	0.9179	0.8625
F1 Score	0.8024	0.7916

Now

- For accuracy, random forest has an accuracy of 74.66% slightly higher than logistic regression 74.55%, meaning it has correctly predicted whether an autopsy was performed or not for approximately 74.66% of the cases in the dataset.
- Precision for random forest is 0.71 & for logistic regression, it is slightly higher 0.73 for predicting autospies as "Yes" or the one labeled as "Y". Therefore the logistic regression model correctly predicted an autopsy 73.16% of the time when it said one occurred, whereas for random forest it is 71.28%.
- For Recall, Random Forest did a pretty good job by correctly identifying an autopsy being performed in approximately 91.80% of the cases where an autopsy was actually performed, on the other hand logistic regression recall came around 86.25%
- For F1 Score, Random Forest (80.25%), which suggests a good balance between its precision(71.28%) and recall(91.80%). Whereas for Logistic Regression(79.17%) also indicating a good balance between precision (73.16%) and recall (86.25%). The Random Forest model seems to perform slightly better in terms of both recall and F1-score, while the Logistic Regression has a slightly better precision.

c. Give 3 examples of individuals that the model predicted correctly. Give 3 examples of individuals that the model predicted incorrectly. Do you notice any similarities within each set of individuals?

   i. Correctly Predicted Individuals:

```
Correct prediction n°1:
The subject died in February of 2015, on a Tuesday, was Female and aged 1 - 4 years.
They were INTERSTATE NONRESIDENTS, their specified race was White, they were Never married, single.
The highest level of education they reached was 8th grade or less.
The death occurred at Hospital, Clinic or Medical Center and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: No
The death was Homicide, the indicated cause of death was: Assault (homicide) (*U01-*U02,X85-Y09,Y87.1)


The Random Forest model predicted that an autopsy was performed, which is correct.


Correct prediction n°2:
The subject died in February of 2015, on a Thursday, was Male and aged 55 - 64 years.
They were RESIDENTS, their specified race was Black, they were Married.
The highest level of education they reached was 9 - 12th grade, no diploma.
The death occurred at Hospital, clinic or Medical Center and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: No
The death was Accident, the indicated cause of death was: All other and unspecified accidents and adverse effects (V01,V05-V06,


The Random Forest model predicted that an autopsy was performed, which is correct.


Correct prediction n°3:
The subject died in May of 2015, on a Friday, was Male and aged 55 - 64 years.
They were FOREIGN RESIDENTS, their specified race was White, they were Marital Status unknown.
The highest level of education they reached was Unknown.
The death occurred at Other and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: No
The death was Suicide, the indicated cause of death was: Intentional self-harm (suicide) (*U03,X60-X84,Y87.0)


The Random Forest model predicted that an autopsy was performed, which is correct.
```

ii. Incorrectly Predicted Individuals:

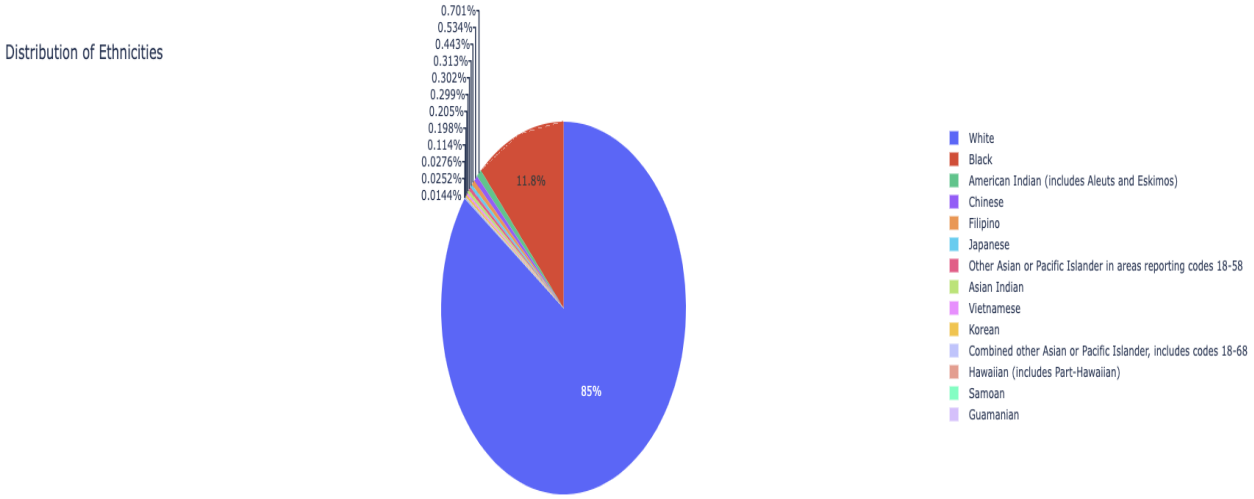
```
Incorrect prediction n°1:
The subject died in July of 2015, on a Saturday, was Male and aged 55 – 64 years.
They were RESIDENTS, their specified race was White, they were Divorced.
The highest level of education they reached was high school graduate or GED completed.
The death occurred at Decedent’s home and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: No
The death was Accident, the indicated cause of death was: All other and unspecified accidents and adverse effects (V01,V05–V06,


The Random Forest model predicted that an autopsy was performed, which is incorrect.


Incorrect prediction n°2:
The subject died in October of 2015, on a Saturday, was Male and aged 55 – 64 years.
They were RESIDENTS, their specified race was White, they were Married.
The highest level of education they reached was some college credit, but no degree.
The death occurred at Other and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: No
The death was Accident, the indicated cause of death was: All other and unspecified accidents and adverse effects (V01,V05–V06,


The Random Forest model predicted that an autopsy was performed, which is incorrect.


Incorrect prediction n°3:
The subject died in February of 2015, on a Monday, was Female and aged 55 – 64 years.
They were RESIDENTS, their specified race was White, they were Divorced.
The highest level of education they reached was 9 – 12th grade, no diploma.
The death occurred at Decedent’s home and the decedent's activity was: During unspecified activity
The decedent was at work at the time of death: Unknown
The death was Could not determine, the indicated cause of death was: All other external causes (Y10–Y36,Y87.2,Y89)


The Random Forest model predicted that an autopsy was performed, which is incorrect.
```

iii. Similarities within each set of individuals

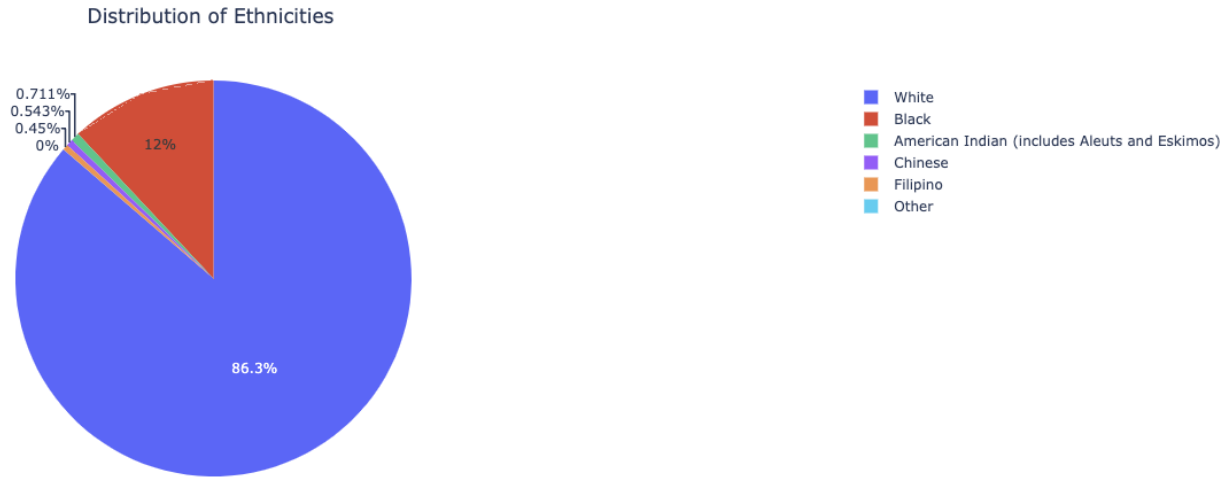
1. Correct Predictions:
    - All the three correct predictions involved cases where an autopsy was performed. for Correct prediction n°1: the person comes under no schooling category.
    - The deaths in the correct predictions were due to homicide, an accident, and suicide.
    - The subjects varied in their demographics: ranging from a young child to males aged 55 - 64 years.
2. Incorrect Predictions:
    - All three incorrect predictions involved the model predicting that an autopsy was performed when it wasn't.
    - The deaths were all classified as accidents or undetermined causes.
    - Most of the subjects were males aged 55 - 64 years, and the deaths occurred in various locations, including the decedent's home.

3. Multi-Class Classification

a. Pie chart for distribution of the different ethnicities



Another clear & updated pie chart is as below



I have created a threshold of 10,000, meaning ethnicities having counts below 10000 will be grouped together into an “Other” category.

b. Clf is RandomForestClassifier & clf3 is logistic regression. Below is the summary of all the model evaluation metrics.

	RandomForest	Logistic Regression
Accuracy	0.8492	0.8502
Precision	0.7823	0.7878
Recall	0.8492	0.8502

F1-Score	0.7884	0.7890

Here the Logistic Regression model (clf2) performs slightly better than the RandomForest model (clf).

- Accuracy: Both models have almost the same accuracy, with the Logistic Regression having an edge by 0.1%.
- Precision: LR, slightly outperforms the RandomForest model in precision by about 0.55%.
- Recall: Identical for both models, which means both models have the same ability to correctly identify the true positive rate.
- F1-Score: Lr has a slight advantage with about 0.06% higher F1-score than the RandomForest.

We can say that the differences are marginal, the Logistic Regression model seems to be the better performer. This slight edge is due to linear behavior of lr. Both models are good, given the marginal difference.

c. Considering pie chart and model performance, population bias - as 85% are white, if the model was primarily trained on or influenced by this demographic, there's a concern that the models might be performing well primarily on the majority demographic and might not be as effective for minority groups. If the majority class is predominantly predicted correctly, while the minority classes are often misclassified, the accuracy would still be high due to the imbalance in class distribution.

   i. The demographic distribution, high precision and recall might be driven by correct predictions in the majority class. Since precision and recall are very close in value for both models clearly showing a balanced performance between false +ves and false -ves.

      There's a potential concern about how well these models generalize across different demographic groups.

Problem 2. Regression

1. Initial Approach
    a. y = df["United States"]

    df_y = y.to_frame(name="Observed_Deaths")

    df_y.head()

```
y = df["United States"]
df_y = y.to_frame(name="Observed_Deaths")
df_y.head()
# --- YOUR CODE ENDS ---
```

Week Ending Date	Observed_Deaths
2017-01-07	59901
2017-01-14	61118
2017-01-21	59445
2017-01-28	58178
2017-02-04	58541

    b. Plot y. Do you notice any seasonal patterns in the number of deaths over time?

- We can see peak in deaths during Covid, and the seasonality varies as per the different covid waves during 2020-early 2022.
- All in All, the cyclical pattern can be due to seasonal illnesses. For example, during winter months, there's an increase in influenza and other respiratory diseases, leading to higher numbers of deaths, especially among vulnerable populations like the elderly, this can be a yearly cyclical

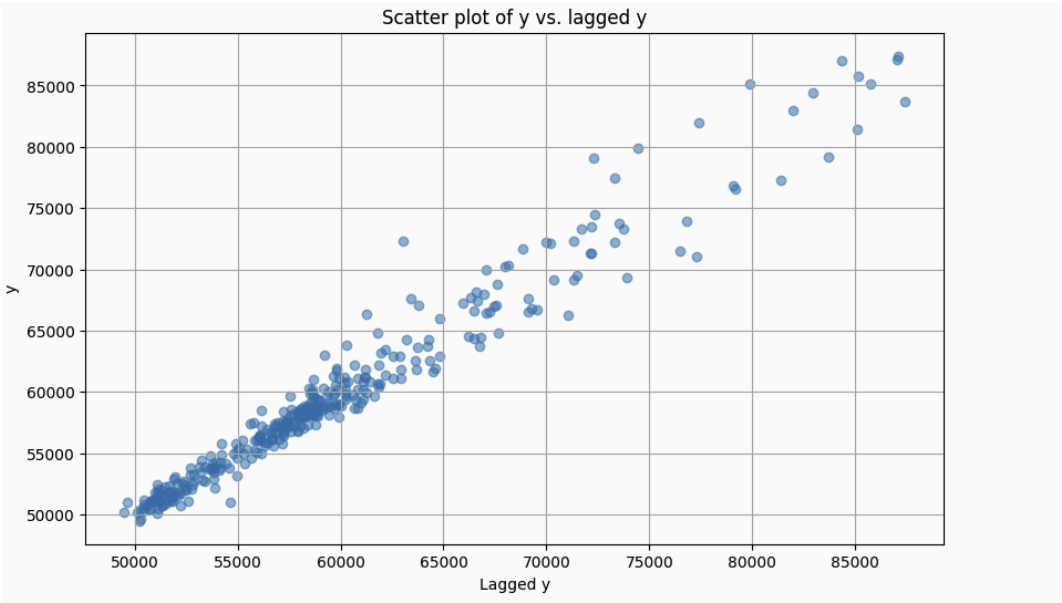
pattern. Therefore from the above statement, we can say during winters there are more deaths.

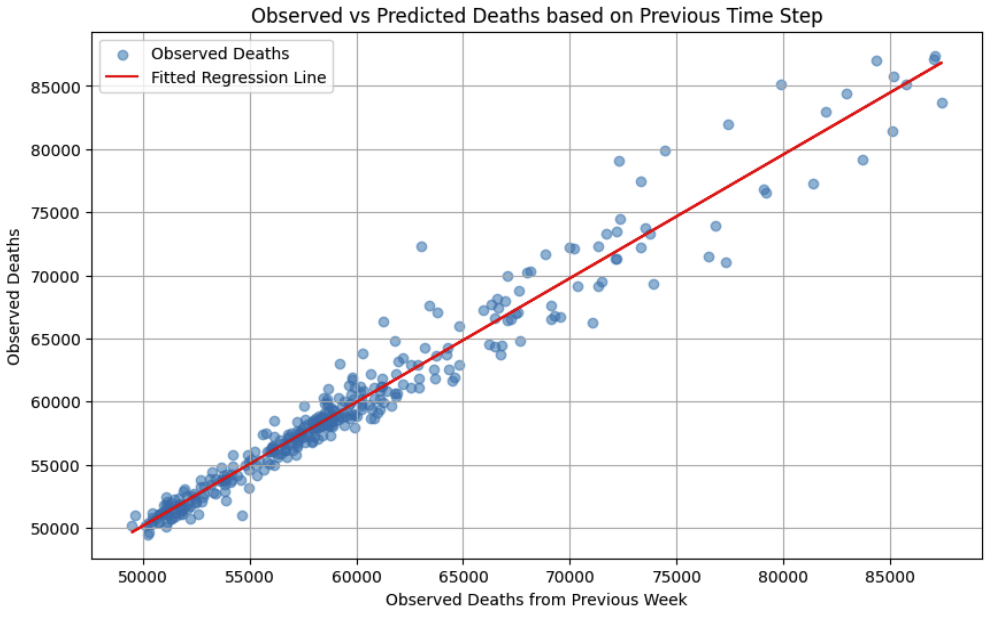
- Seasonal factors like heat waves in summer or extreme cold in winter
- Traveling or gatherings during holidays leading to more accidents or disease spread.
- We can also see that there is overall increasing cyclical pattern, and also decreasing - this might be due to increasing population (for hikes) and healthcare improvements (for decreasing trends).

c. reg = LinearRegression()

reg.fit(x, y)

d. Scatter Plot





Fairly accurately predict the number of deaths in a week based on the number of deaths in the previous week. We can say that there is a strong linear relationship between the number of deaths in a week and the number of deaths in the preceding week or the number of deaths in one week is largely influenced by the number of deaths in the week before.

e. Adjusted R2 coefficient: 0.9579311494948856 .

   95.8% - strong goodness of fit. Approx 95.8% of the variance in the observed deaths (y) can be explained by the model using the previous week's observed deaths.

2. Can we develop a better model?
   a. reg2 = LinearRegression()

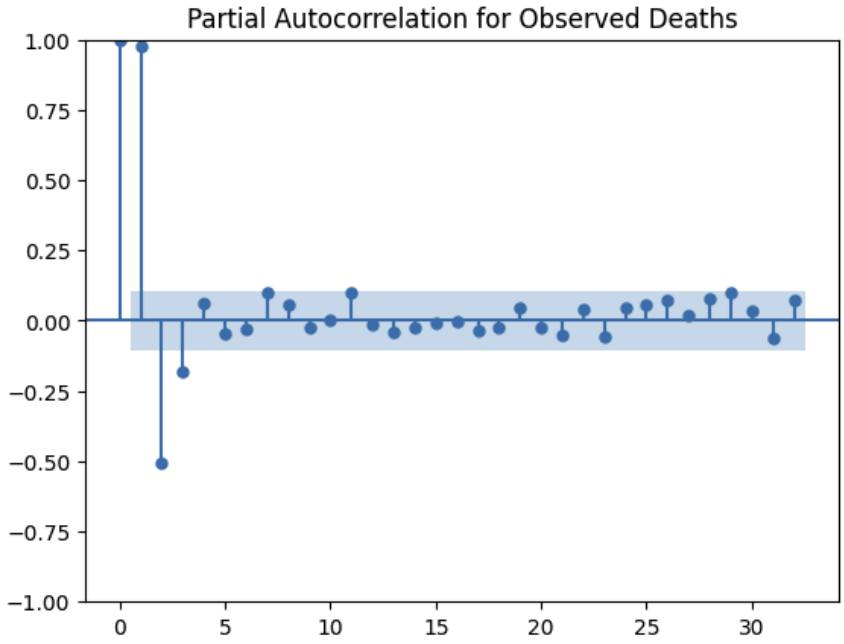
      reg2.fit(x, y)

```
# 2.a)
# Q: As in the previous part, build an AR(2) model using linear regression.
# Shift the data to create aligned x and y variables
x = pd.concat([df["United States"].shift(2), df["United States"].shift(1)],
              axis=1)[2:].to_numpy()
y = df["United States"][2:].to_numpy()

# Create a LinearRegression model and store it in a variable named reg2
# --- YOUR CODE HERE ---
reg2 = LinearRegression()
# --- YOUR CODE ENDS ---

# How do you fit reg2 to x and y?
# --- YOUR CODE HERE ---
reg2.fit(x, y)
# --- YOUR CODE ENDS ---
```

```
▾ LinearRegression
LinearRegression()
```

   b. Adjusted R2 coefficient: 0.9723444330032064, 97.23% strong goodness fit.
      - The value is now closer to 1, defining strong explanatory power.
      - AR(1) model had 0.9579 and AR(2) now has 0.9723.
      - AR(2) captures more of the variance than AR(1) model.
      - More lag could increase the fit, but there can also be a risk of overfitting.
      - We can consider techniques like cross validation to validate model performance.

c. The partial autocorrelation function



(Lag longer than 2 weeks) or more lag could increase the fit, but there can also be a risk of overfitting.

d. Output is as follows:

   i. According to our prediction, at 2023-09-23 00:00:00, there will be 49157.84331430184 deaths during the week.

      No of deaths forecasted is 49157 deaths & the date is 23rd Sept 2023.