United States Death Trends

**Introduction**

Study and analysis of Mortality and its causes is critical for the interpretation of national trends and international comparisons. Mortality statistics are a mirror to the situation of health prevalent in a population. The mortality trends depict the picture of population’s health status. In depth analysis of deaths and its indicators can substantiate the progress of health programs

**Problem:**

The life expectancy of the United States is in decline. What are the leading causes for the deaths of Americans?  Finding the major causes and change in epidemiological outbreaks can help mitigate some of the major death leading causes.

**Data Set**

<https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu>

The dataset that is being used is taken from United States Center of disease control and prevention (CDC) from 1990 to 2017. This study analyzes the leading causes of deaths in United States of America between 1999 and 2017. Dataset were downloaded as CSV file.

**Data Preprocessing and Cleaning:**

Pandas read\_csv() is used to read the file. After checking df.info and dtypes, df.head() gave a preview to the first 5 rows of dataset. Using shape function on dataset we see approximately 10868 cases of death cases were recorded in different US states. Using python built in function null values were searched. There were no null values in our dataset. Datetime is designated as object which is converted to python date time using pandas datetime (pd.to\_datetime) function.

113 Cause Name column is splited into other two column cause 1 and cause 2. Cause 2 column is dropped whereas cause1 column is renamed to disease cause. Column ‘113 cause name’ is dropped.

Data for United States were only used for further analysis. Column named as ‘State’ is renamed to ‘Country’ column.

Unique Death Causes in the United States were searched using unique function. There are approximately 11 unique causes excluding row containing ‘All’ causes.

**Exploratory Data Analysis**

With the help of exploratory data analysis, I tried to answer some of the questions which are listed below:

• What is the total number of records in the dataset?

• What were the causes of death in this data set?

• What is the number of deaths per each year from 1999 to 2017?

• What were the top causes of deaths in the United States during this period?

• 10806 no of cases were recorded.

• Accidents (unintentional injuries)', "Alzheimer's disease", 'Cerebrovascular diseases', 'Chronic lower respiratory diseases', 'Diabetes mellitus', 'Diseases of heart', 'Influenza and pneumonia', 'Intentional self-harm (suicide)', 'Malignant neoplasms', 'Nephritis, nephrotic syndrome and nephrosis'. Unique cases responsible for number of deaths over a period of year 1999- 2017 is depicted on figure 1.

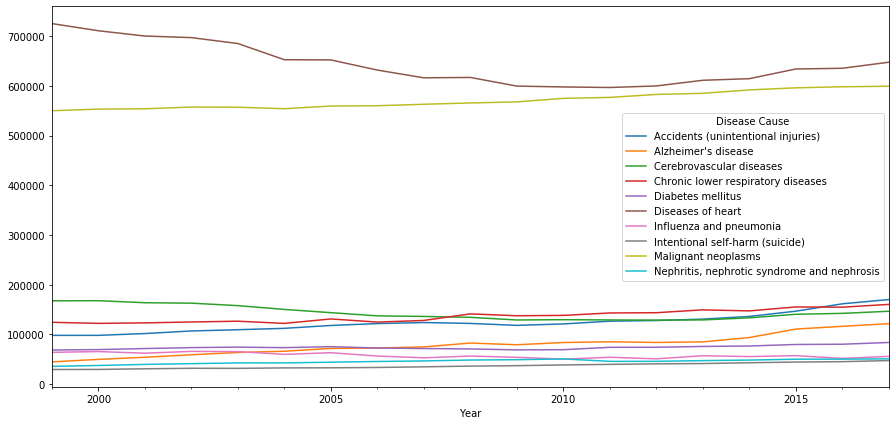


Figure 1:-Causes of Deaths Over a Period of 1999-2017

• The number of deaths declined between 2002 and 2009. Then there was a continuous growth in the number of deaths from 2010 to 2013. Finally, there was a sharp increase in the number of deaths in 2013 and 2014 which can be explained by figure 2.

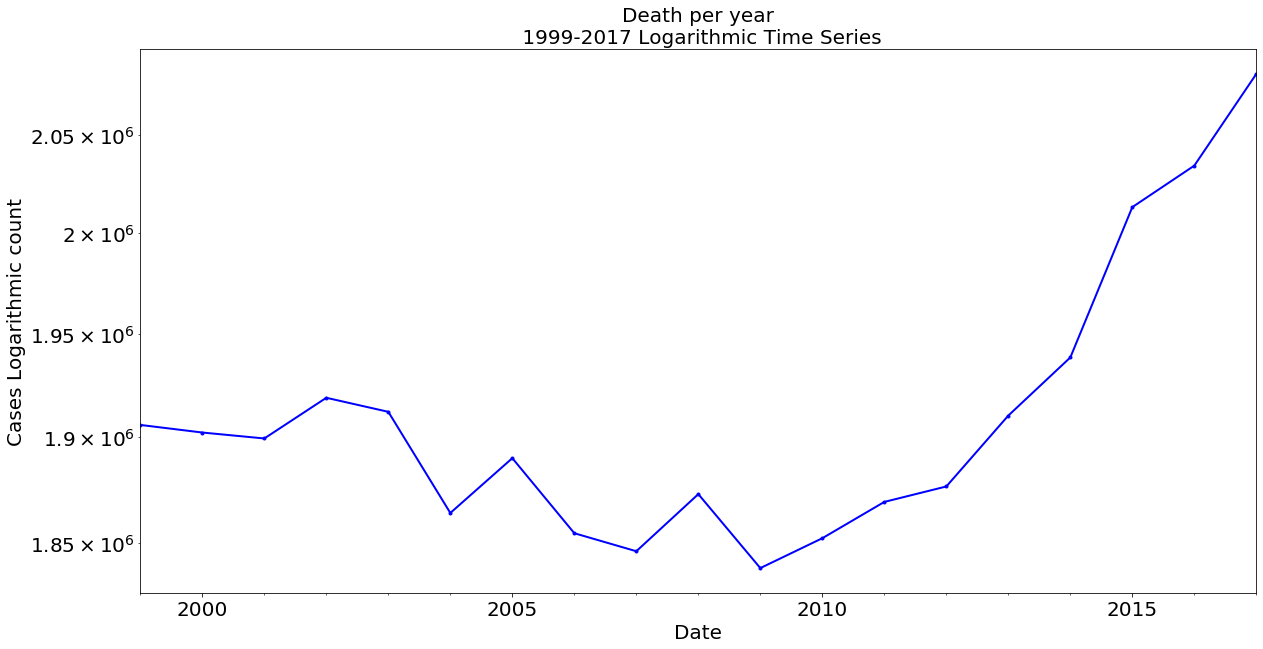


Figure 2 Number of Deaths over the period of 1999-2017

• Diseases of the heart represent the highest causes of death followed by cancer which is visualized on figure 3.

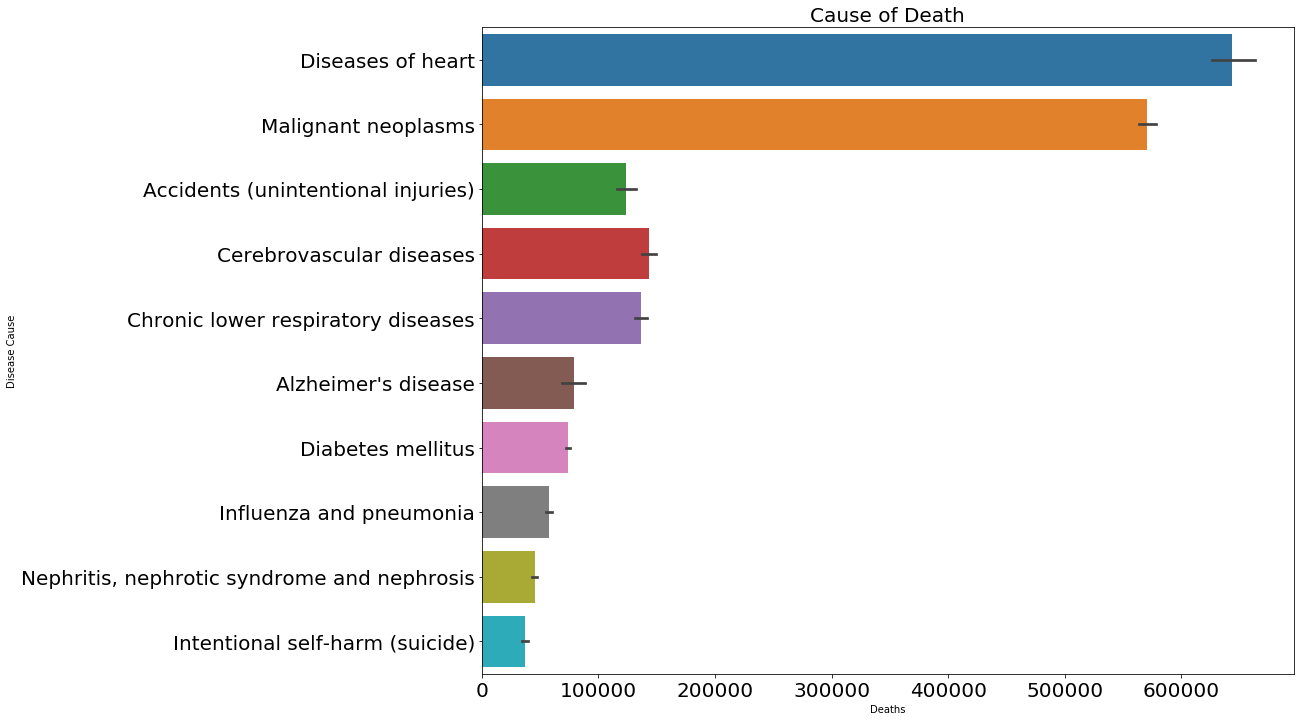


Figure 3 Prevalence Unique Cause of Deaths

Exploratory data analysis of our data shows no of deaths are increasing and the major contributing causes belongs to non-communicable disease like heart disease and cancer.

# **Statistical Analysis of Capstone Project One**

The purpose of statistical analysis of our dataset is to explore some of the basic concepts of statistics using capstone project one dataset. We started using basic python condes to explore the descriptive statistics of dataset. Dataset has 10868 observations in all columns. Same number of observations also implies that there is no missing values in our dataset.

**Ordinary Least- Squares (OLS)** Regression technique is used for statistical learning of the dataset using the statsmodel python package. OLS uses squared error which has nice mathematical properties thereby making it easier to differentiate and compute gradient descent. This method is easy to analyze than more sophisticated models and computationally faster. We used Age-Adjusted Deaths as a dependent variable (y) and other attributes year, deaths, causes, log of deaths as independent variable(X). Age adjusted death rate is a measure that controls for the effects of age differences on health event rates. When comparing across geographic areas, age adjusting methods is used for the influence that population age distributions might have an health even rates (<https://ibis.health.state.nm.us/resource/AARate.html>.)

**Checking the OLS Assumptions**:

**Hypothesis Testing**: Since this is a Multiple Linear Regression, we need to know the importance of variables(significance) with respect to the hypothesis. To do this, we need to calculate the p value for each variable and if it is less than the desired cutoff (0.05 is the general cut off for 95% significance) then we can say with confidence that a variable is significant. Our OLS model shows all variables are significant as the p value is less than 0.05

**R Square** (Coefficient of Determination): - This metric explains the percentage of variance explained by covariates in the model. It ranges from 0-1. R square calculated for our model is 0.50 (approx.) meaning 50 percent variability of attributes of our sample.

**F Statistics** - It evaluates the overall significance of the model. It is the ratio of explained variance by the model by unexplained variance. It compares the full model with an intercept only (no predictors) model. Its value can range between zero and any arbitrary large number. Naturally, higher the F statistics, better the model. Our model has f statistics of 2129 which implies that overall regression is meaningful.

**Checking for multicollinearity for regression**: - The variance inflation factor is used to check multicollinearity on our model. Vif value equal to 1 no multicollinearity. Table 1 shows the value of Vif .Vif value equal to 1 show no multicollinearity.

|  | **VIF** | **Features** |
| --- | --- | --- |
| **0** | 1.325250 | Year |
| **1** | 1.076213 | Deaths |
| **2** | 1.399952 | Age-adjusted Death Rate |

Figure 4 VIF of Features

**Residual vs. Fitted Values Plot**

Ideally, this plot should not show any pattern. But if you see any shape (curve, U shape), it suggests non-linearity in the data set. In addition, if you see a funnel shape pattern, it suggests your data is suffering from heteroskedasticity, i.e. the error terms have non-constant variance



Figure 5 Residual Vs Fitted Curve

**Normality Q-Q Plot**

As the name suggests, this plot is used to determine the normal distribution of errors. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then your residuals have a non-normal distribution (problematic situation).

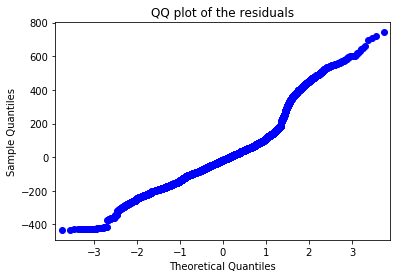


Figure 6 Q-Q Plot

References:

* <https://ibis.health.state.nm.us/resource/AARate.html>.)
* <https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html>
* <https://statisticalhorizons.com/multicollinearity>

# **Application of Machine Learning to Age adjusted death rate modeling and forecasting**.

The purpose of this study is to investigate a prediction model that identifies the potential age adjusted mortality rate in United states using machine learning based on leading causes of deaths data collected by the CDC United States from 1999 to 2017. The data details 11 unique causes of deaths and includes states, year, and number of deaths caused by unique causes.

Multiple linear Regression

Random Forest Regression

Xgboost Regression

Transform the Input Data

Pre-Process the Data

Obtain Dataset for Leading Causes of Deaths

Select the best model after hyperparameter tuning

Metric: Accuracy

Metric: RMSE

Figure 1 Diagramatic Approach for the workflow of methodology

**Test and Training Data** The training dataset represents 67% of the original data which was used to build the model; the test dataset represents the remaining 33% of the original data which was used to assess model performance.

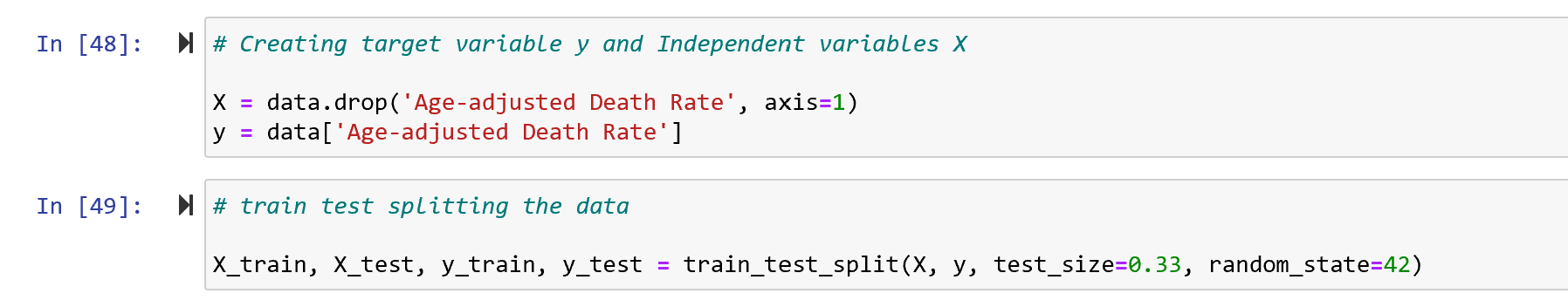


Figure 2 Train Test Splitting data

**Multiple Linear Regression**: Multiple regression analysis was used to determine the relationships between the dependent variables Y to discover which relationships were linear. The following equation shows the relationships between Y and X1, X2, . . ., Xn

Υ = β0 + β1X1 + β2X2 + . . .. +βnXn + e

where Y is the dependent variable, X1–Xn are the independent variables, β0 is a constant (or y-intercept), and β1– βn are the coefficients of the respective variables (loading or partial slopes), which are used to describe the change in the Y value when the X value changes. RMSE calculated for this model was 161.

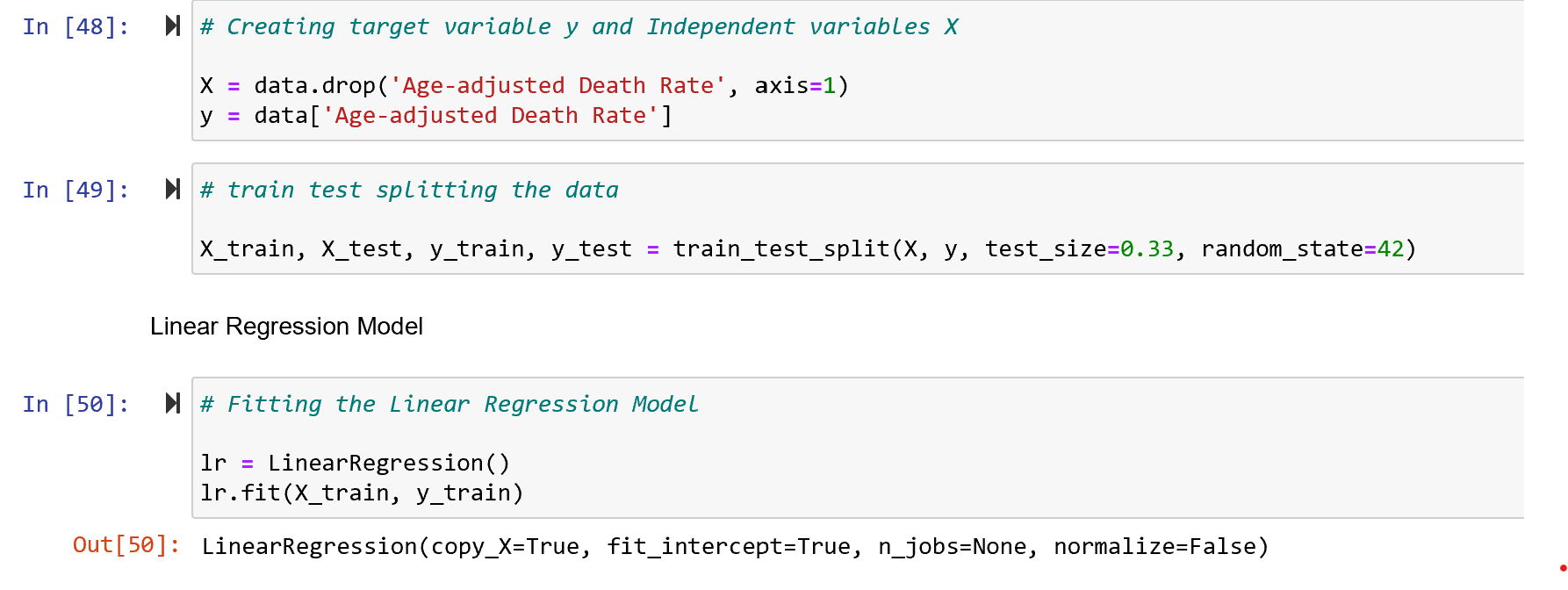
****

Figure 3 Linear Regression

**Random Forest Regression**: - Random forest (RF) is a supervised decision tree algorithm that bags un-pruned trees by using randomly selected covariates at each split. The accuracy for test dataset from this model was 99% and RMSE was found to be 12.

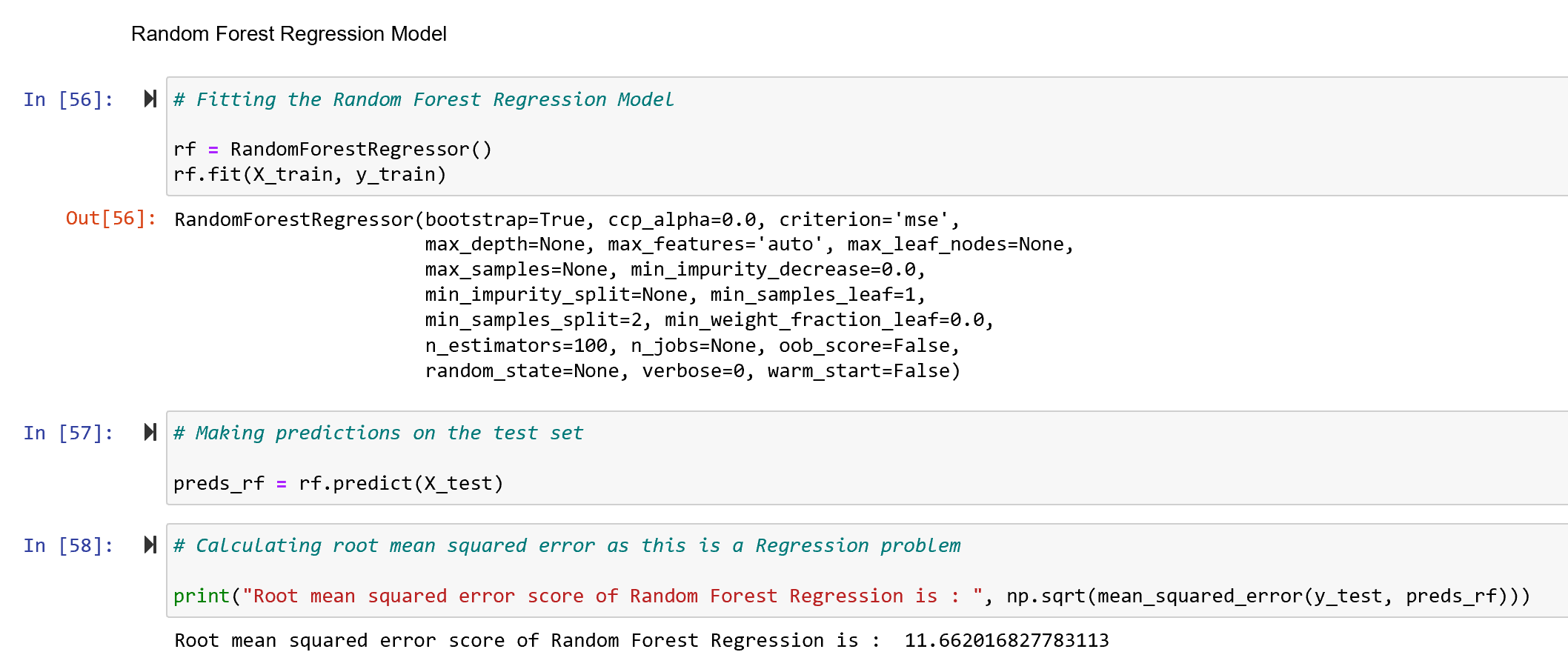


Figure 4 Random Forest Regression

**Xgboost Regression: -** Gradient Boosting (Xgboost) is a successful machine learning library based on a gradient boosting algorithm proposed by Tianqi Chen. It has better control against overfitting by using more regularized model formalization, in comparison to prior algorithms. The Xgboost model was fitted using GridSearch and RMSE was calculated which shows decline in RMSE after hypertuning the parameters for the model.

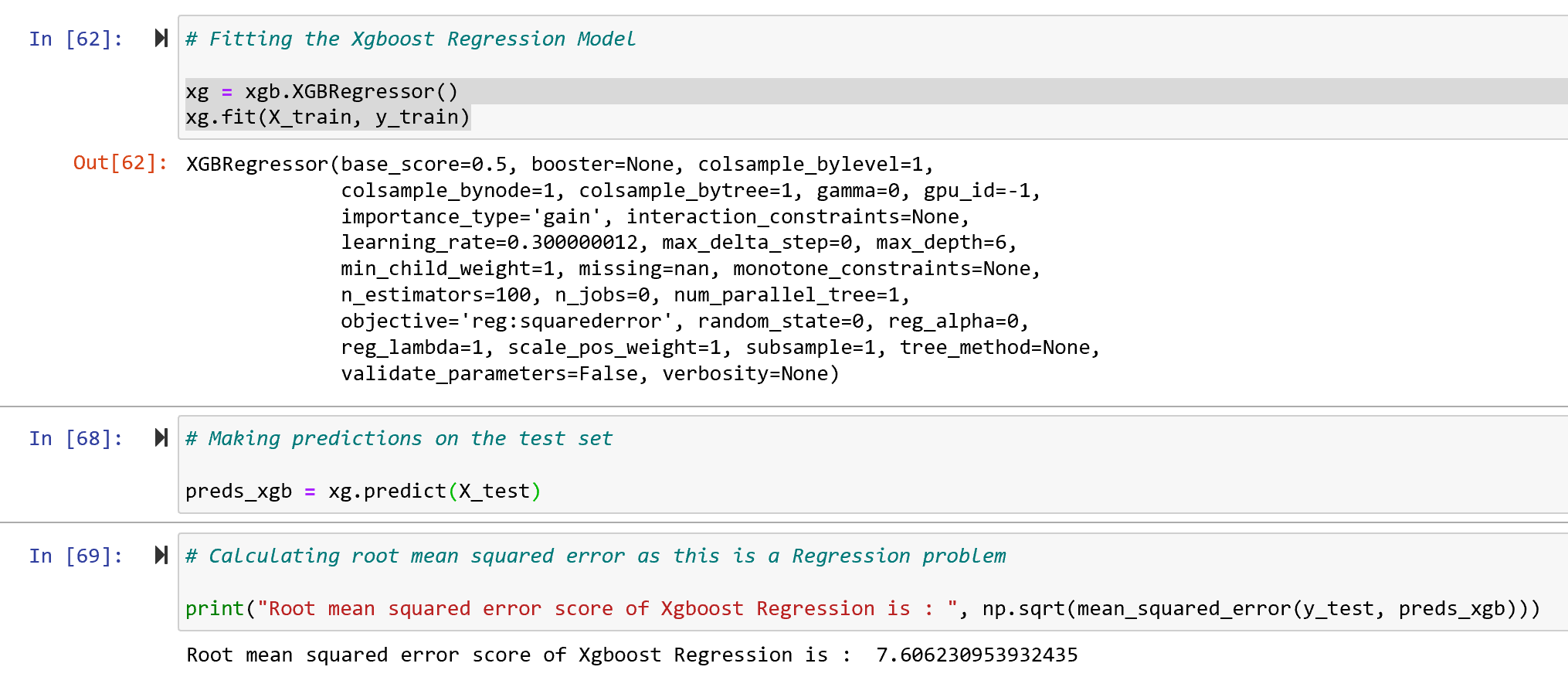


Figure 5 Xgboost Regression

**Root Mean Square Error (RMSE): -** RMSE is the most common statistical metric of a regression model to measure the performance of model.

**Fine tune your model:** -After calculating RMSE for all the models, model with low RMSE was fine-tuned using Scikit learns GridSearch.

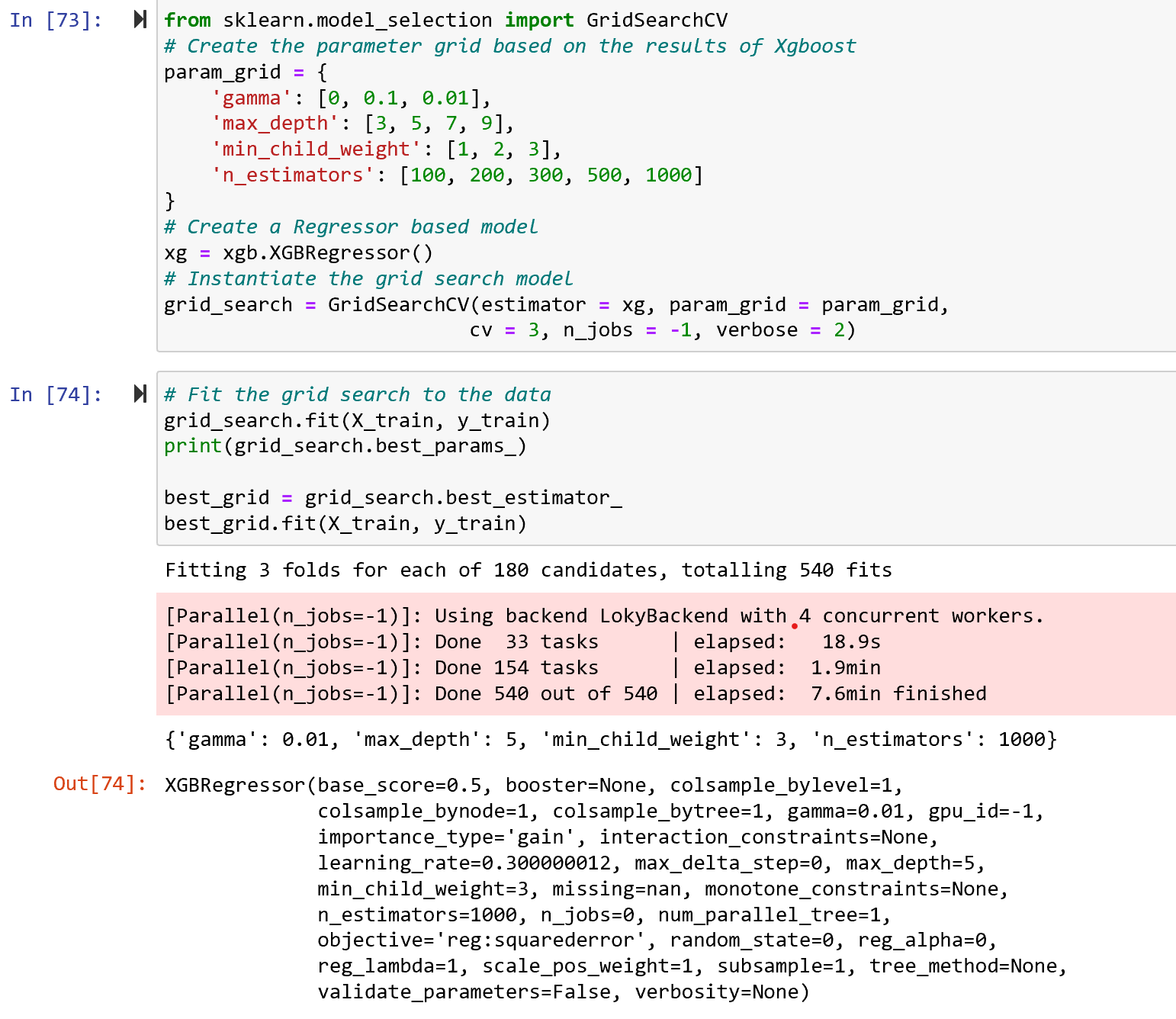


Figure 6 Hypertuning of Parameters

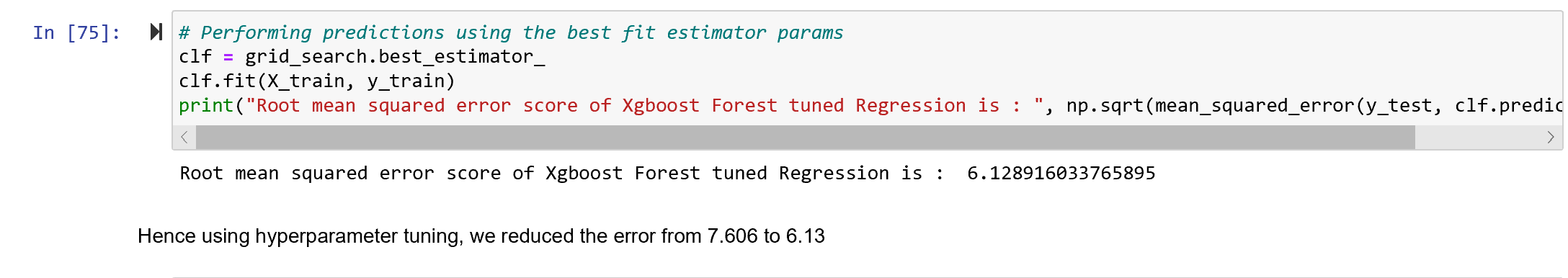


Figure 7 RMSE of Xgboost Regression after Hypetuning

**Findings: -** Upon describing the distribution of the dataset, machine learning methods, such as linear regression, random forest regression, and XgBoost analyses were applied with the derivation of major variables influencing classification in each algorithm. A comparison of the performance of each model showed the root mean square error to be lowest for the Xgboost method, at 6.17, which translates to a 99.9% successful predictive rate in terms of classifying age adjusted death rate. To quantify the usefulness of all the variables in the entire random forest, we can look at the relative importance’s of the variables. The random forest analysis of this study indicates that the cause of deaths are the most influential factors, followed by deaths and state based on feature prediction importance graph shown in figure 2.



Figure 8 Feature Importance

Hence, this study demonstrates the feasibility of machine learning in the predicting age adjusted mortality rate on different states based on leading causes of disease. The results obtained can contribute to the prevention of leading cause of deaths by raising awareness of potential risks, by quantitatively predicting age adjusted fatality and incorporating the findings with healthcare industry.

**References:**

<https://ibis.health.state.nm.us/resource/AARate.html>.)

<https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html>

<https://statisticalhorizons.com/multicollinearity>

Muller A. Education, income inequality, and mortality: a multiple regression analysis. *BMJ*. 2002;324(7328):23‐25. doi:10.1136/bmj.324.7328.23

Yu W: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making. 2010, 10 (1): 16-10.1186/1472-6947-10-16.

Breiman L: Random forests. Machine learning. 2001, 45 (1): 5-32. 10.1023/A:1010933404324.

Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD ’16, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794