A Dissertation on

PREDICTING THE BUSINESS IMPACT OF COVID-19

Submitted to
University of Nottingham

In partial fulfilment of the requirements for the award of the degree of

Master of Science
in
Business Analytics

By

JIGNESH MANOCHA

Under the guidance of

Dr. GEORGIANA NICA-AVRAM
Transitional Assistant Professor

NOTTINGHAM UNIVERSITY BUSINESS SCHOOL

September 2021

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and thanks to my supervisor Dr. Georgiana Nica-Avram who despite being busy with her duties, took time to hear, guide and keep me on correct path to complete this dissertation project.

Besides I am forever thankful to my parents who makes it possible for me to pursue masters from this esteemed university.

Next, I would like to thank my friends and family for the constant support throughout the course.

And at last, my special thanks to all the professor and teachers for guiding me throughout the life.

JIGNESH MANOCHA

Table of Contents

Abstract

The COVID-19 outbreak is the most prominent world crisis since the Second World War. The pandemic that began from Wuhan, China in late 2019 has influenced every country of the world and set off a worldwide financial emergency whose effect is felt and will be felt for quite a long time in future. This requires a need to anticipate pandemic predominance to control the forthcoming impact. This study aims to gain an insight from the descriptive examination and correlation analysis of sociodemographic characteristics of 317 local regions of England. Based on those insights an estimator is developed that would anticipate the impact of coronavirus on the employment furlough level of the businesses. Various statistical techniques and various machine learning algorithms has been implemented. Based on those best results developed on certain evaluation metrics, the chosen model was fitted in the dataset keeping the problem of over and under fitting in mind. The results suggested predication of the employments furlough level and noted a great variation in the pattern of actual and predicted output. The main aim of the study is to address the model to the policymakers so that the future unforeseen circumstance is avoided emerged due to the coronavirus outbreak.

CHAPTER 1

INTRODUCTION

The World has been gripped by a pandemic since the first half of 2020. It was distinguished as a new Covid (extreme intense respiratory condition Covid 2, or SARS-CoV-2), and later named as Coronavirus Disease-19 or COVID-19 (Qiu et al., 2020). The contagious novel coronavirus (COVID-19) was first found in Wuhan, China in December 2019. Since then, the whole world emerged as a new transformed hub of chaos. Human health is massively affected, and the global economy suffered a greatest dip over the century. Alongside it had devastating impact on the global business sector. Many businesses shut their activities, great number of employees got deployed from their work and faced an unprecedented challenge at their work.

Even today, government has failed to successfully predict the upcoming effects of novel coronavirus on the business, as many businesses reported a fall curve of their sustainability. Since artificial and business intelligence are transforming the way of doing business in an automated process, it is vital to perform the data-driven analysis that would measure the impact and forecasts the upcoming calamities and empowers the world in advance to mitigate the impact of virus on population and business as a whole.

To solve the problem of anticipating the impact on business activities, the study has been conducted to perform correlation and regression analysis of the socio demographic elements which impacted the businesses so far, existing in the local areas of England. This study reviews the existing related studies undertaken to assess this problem and establish the relationship between the sociodemographic traits of the areas and self -reported covid cases and mortalities in order to estimate the jobs on furlough.

1.1 Research Aims and Objectives

The study determines two valid objectives to make the analysis to a specific criterion which helped to get a deep insight of the key features established to perform the correlation and gain result out of it.

- To examine the sociodemographic factors contributing to impact.
  - Income
  - Education, Skills and Training
  - Health and Disability
  - Crime
  - Barriers to Housing
  - Living Environment
  - Employment

- To estimate the number of furloughed jobs based on sociodemographic factors.

1.2 Research Questions

To inspect the study that how the risk to covid-19 incited into furlough challenges are dispersed across the dynamic business employment during the initial stage of the outbreak. There was an examination of the insights from different features:

(1) What is the deprivation status and most and least affected areas?

(2) What is the association between the sociodemographic factors and employment furlough levels and the magnitude of their bond?

(3) What is the major infected areas overtime?

The next evaluation details that whether these factors led to build predictive model:

(1) What are the supporting input factors to predict the impact?

The concluding comments will focus on the best way to decipher the reported significant relationships and anticipated furlough considering the temporal context of the pandemic.

CHAPTER 2

LITERATURE REVIEW

In this chapter selected literature pertaining to the impact of covid-19 on businesses, socioeconomic factors determining the effect and methodologies used have been reviewed.

Since the declaration of the presence of the severe acute respiratory syndrome Covid 2 (SARS-CoV-2), which causes the Covid sickness 2019 (COVID-19), the world has suffered huge stress. On December 30, 2019, the Chinese Wuhan local wellbeing authority gave an epidemiological alarm following the main archived case on December 8, which prompted the conclusion of the Huanan fish market two days after the fact (Huang et al., 2020). The World Health Organization (WHO) affirmed the disease as a worldwide pandemic on March 11, 2020 (Cucinotta and Vanelli, 2020).

The Covid-19 has reported in over millions of cases and deaths worldwide. It has ignited the fears of an approaching recession and economic crisis. The drop in the workforce across all economic sectors and unemployment were the key areas affected due to self-isolation, travel restrictions and social distancing (Nicola et al., 2020). The covid-19 pandemic has disturbed each aspect of life universally. Trade and business are crucial areas where the money related crunch has been intensely felt.

2.1 Impact of COVID-19 on Businesses

Businesses have faced different difficulties during the outbreak of pandemic COVID-19, and their reaction towards this disruption has affected their resilience and their chances to beat this crisis. Small and medium-sized enterprises are changing their plan of action to adjust with this evolving environment (Gregurec, Tomičić Furjan and Tomičić-Pupek, 2021). The coronavirus outbreak has shut numerous businesses, leading to an extraordinary interruption of trade in most industry areas. Retailers and brands faced some short-term difficulties, for example the health and safety – its production network, workforce, cash flow, marketing, sales, and customer demand directing these difficulties effectively won't ensure a promising future, because once this pandemic is passed, a different world will emerge opposite to the one preceding the outbreak.

Many business sectors, particularly in the fields of the hotel and travel industry, no longer exist. Renowned US companies like JC Penny, J. Crew, Sears, Hertz, and Neiman Marcus were under huge financial problems. There had been 90% plunge in the airlines' workforce and the travel industry is enormously affected as 80% of rooms in the hotel were vacant (CNN, 2020). All business operations are expected to focus on and improve spending or defer the task that will add less value. Organizations, particularly new startups, have executed an endless employing freeze. Simultaneously, the online shopping and entertainment reported an unprecedented growth (Donthu and Gustafsson, 2020).

The number of active businesses in the United States plunged by the 3.3 million or 22% over the crucial initial 2 months of the pandemic, which is the largest drop in the record and losses to business activity nearly in all industries. African-American businesses reported the bad experience and drop in the business activity by 41%, the business activity of Latinx and Asian business owners were fell by 32% and 26% respectively. Furthermore, it is claimed that these losses to business activity have a severe implication for profit losses, policy making and future economic inequality (Fairlie, 2020).

There had been lower mortality rates for the local municipalities with higher exposure to shut down the sectors using the difference-in-differences approach during the covid-19. The business closures showed quickly avoidable losses and had enormous impacts outside the closed businesses themselves, including spillovers to other local authorities. However, the results reported that the business closures are beneficial when they are selectively executed and coordinated centrally (Bongaerts, Mazzola and Wagner, 2021).

Lockdown and social distancing policies leads to rise in the unemployment in the US (Lozano Rojas et al., 2020). There had been plunge in the employment rate in the US by 1.7% for each additional 10 days that a state encountered a stay-at-home command during the period Mar 12 - April 12 (Gupta et al., 2020).

(Coibion, Gorodnichenko and Weber, 2020) computed a serious fall in the workers participation rate over the long haul along with the rise in the "demotivated workers" (unemployed labours who have effectively quit looking for work viably pulling out from the labour force). This may be because of the unbalanced effect of COVID-19 on the older population. (Aum, Lee and Shin, 2020) identified that an increment in contaminations results

to a drop in local employment without lockdown restrictions in South Korea, where there were no administration's ordered lockdowns. This number expanded for nations, for example, the US and the UK where compulsory lockdown measures were forced.

2.2 Case Study – Somaliland

In Somaliland, the COVID-19 outbreak has shut numerous businesses and stores, leading to an unpresented disturbance of business in most industry areas. Retailers and brands faced some short-term difficulties, for example the health and safety – its production network, workforce, cash flow, marketing, *(*Donthu and Gustafsson, 2020). Cost of bills has been increased by private power companies because of their dependency on fuel imports from abroad countries. Hence, price changes in all administrations and services have significant impact on all Somaliland's business areas (Mohamed, 2021). The pandemic has massively impacted the Somaliland's small businesses and around 25% of them were shut down (Bartik et al., 2020).

The COVID-19 is probably going to cause insolvency for some renowned brands in numerous industries as buyers stay at home and economies shut down (McKee and Stuckler, 2020). This isn't just having impact for the economy; all of society is influenced, which has resulted in drastic change in the consumer and businesses behaviour (Donthu and Gustafsson, 2020). Most significant ventures confronted huge drops in the number of business owners with the exceptional case of agriculture. Construction, cafés, transportation, and hotels all confronted enormous drop in the businesses' number because of COVID-19 (Fairlie, 2020).

As per International Labour Organizations, 2020, the pandemic outbreak has already put the labor and economic market into shock, affected demand, supply, investment and consumption (International Labour Organization, 2020). Nonetheless, major business and economic activities have been affected by the restrictions and lockdowns inflicted by the public authority of Somaliland, resulted in rise of unemployment level, for instance the travel restrictions and social distancing measures leads to closure of sports and amusement centres, both private and public colleges and schools. Employers have been set under a serious pressure test as interest for their services soar in the initial not many months of the outbreak while their capacity was seriously obliged by the need to reduce close and personal contact with job seekers (OECD, 2020).

The analysis on implications of covid-19 on small businesses using the survey data of 5800 small scale businesses (members of Alignable) have reported that 43% of the businesses were temporarily shut off and employees count have been plummeted by 40%, most of the businesses were financially struggling in meeting the monthly expenses and maintaining liquid assets and majority of businesses were seeking fund from the local authorities (Bartik et al., 2020).

The study on the changes in entities' activities to measure the business performance in response to the COVID 19 pandemic using dataset of 218 Romanian listed different scale companies reported the 37.43% drop in the net profit of overall market in contrast to companies engaged in commerce, IT, agriculture, transport, R&D which reported the improved financial performance. The results suggest that effective liquidity management, an increases company size and equity financing consolidate the economic performance of entities related to the equity and asset's return (Achim et al., 2021).

2.3 Insights from existing methodologies

(Lee and Chen, 2020) employed a quantile regression model as a proxy of OLS approach in order to perform analysis on the impact of covid-19 on the travel and leisure industry returns as QR estimates the reactions of output variable at assorted quantiles by offering the proof beyond the average of the data. Following the methodology, it was identified that the V-shape correlation between the travel and leisure industry returns and the number of cases recovered across return quantile and the positive correlation between government response stringency index and returns.

The research on the different socio-economic determinants of global covid 19 mortalities (Ngepah, 2021), used the Poisson pseudo-maximum-likelihood (PPML) and the quantile regression method to exploit the non-linear estimates of the data to make the interquartile comparisons so that it can be used to recommend which societal features become crucial at disastrous level when existing healthcare systems become overwhelmed. (Ehlert, 2021) The socioeconomic, demographic and health related determinants of covid 19 at the regional level cases and deaths in Germany. A multivariate spatial model has been used that include 401 countries in Germany to report for regional interrelations and spillover impacts. It was seen

that the spillover effects on the total cases of neighbouring regions were identified for specific factors, with a different sign in comparison to the general impacts.

The analysis of socioeconomic determinants of COVID-19 infections and mortality was undertaken considering the region of England and Wales using the correlation and regression analysis which shows a particular relation between the number of confirmed covid 19 cases and the number of deaths with covid 19 per 100,000 people. The report reveals the local areas with larger households, worse level of self-reported health and large amount of folks using public transport are majorly infected by the coronavirus. Moreover, there is a weak correlation in mortality, household size and use of public transport and strong correlation at the old age population (age), black and Asian citizens (ethnicity) and self-reported health which means these are the most reported features for the death due to covid 19 (Sá, 2020).

Various models have been utilized in late examinations to foresee impact, predominance, and death rate of COVID-19. (Li, Feng and Quan, 2020) For example, the strategy to anticipate the outbreak size and ongoing trend in China using data driven analysis. (Wang et al., 2020) There was an open dataset to assess the destruction rate of Coronavirus outbreak with the help of Patient Information Based Algorithm. (Roda et al., 2020) A correlation analysis on the standard SIR and SEIR structures to model COVID-19 in Wuhan, China.

Statistical and time series model has been employed to build model and anticipate the prevalence of this pandemic. (Ghosal et al., 2020) The usage of linear regression model to forecast the mortalities in India because of SARS-CoV-2. (Ceylan, 2020) Implementation of Auto Regressive Integrated Moving Average (ARIMA) model to anticipate the pervasiveness of COVID-19 in Italy, Spain, and France.

Machine learning algorithms have been used to estimate the potential outputs and in numerous applications which requires the anticipation of unfavourable risk factors. (Mojjada et al., 2020) The study to anticipate the effect of covid-19 as a threat on individuals using ML modelling.

Four regression models have been used in the study to predict the impact of covid -19

- Linear Regression
- LASSO Regression
- Support Vector Machine
- Exponential Smoothing

Machine learning can be used to extract valuable data from huge datasets and build robust data-driven prediction models for medical care. (Rustam et al., 2020) It was anticipated that the risk factors behind spread in covid-19 outbreak using various machine learning algorithms for example, Exponential Smoothing (ES), least absolute shrinkage and selection operator (LASSO), LR and SVM. (Sujath, Chatterjee and Hassanien, 2020) Implementation of various machine learning prediction models, for instance multilayer perceptron (MLP), vector autoregression and linear regression (LR) algorithms to predict the cases in India using Covid-19 dataset from Kaggle. To forecast of the covid-19 cases across different countries (Yadav et al) different machine learning models were used such as naïve Bayes (NB), decision tree (DT), support vector machine (SVM), Linear regression, random forest (RF) etc.

Machine learning models such as SVM (linear and RBF), KNN, Lasso and Random Forest has been developed to anticipate the mortality of covid-19 patients. The algorithms used were trained to recognize three cases, i.e., mortality and survived within 14 and 30 days period after the initial diagnosis, survived and mortality. Linear SVM proved to be the winning model with a sensitivity of 0.92, specificity of 0.91 and AUC of 0.962. The investigation found age, mellitus, diabetes, and cancer is the critical factor in the mortality prediction (An et al., 2020).

2.4 Evaluation Metrics

Root mean square error (RMSE) is the most broadly used performance evaluation metric to estimate the regression algorithms. Different measurements, for instance R2-Squared and Mean Absolute Error (MAE), are used alongside RMSE (Satu et al., 2021).
The predictive models are evaluated utilizing two statistical metrics: Root mean square error (RMSE) and determination coefficient. These indices give insights into the algorithm integrity and accuracy fit for the dataset (Saba and Elsheikh, 2020).

Regression Analysis has been employed to examine the relationship between independent and dependent variable in order to analyse the degree of impact of independent variables (IMD domain, covid cases and deaths) on the dependent variable i.e, employment furlough level.

CHAPTER 3

RESEARCH AND METHODOLOGY

This chapter of the study addresses the research aims & objectives. It also highlights the approaches, concepts and the algorithms used therein.

- **Research Objective 1**

  The first and foremost objective of this study explains the relationship among socio-demographic factors which impacts the businesses.

  The aim was to gain an understanding of the variables' behaviour and their significance with each other. The main expected outcome of this objective was the descriptive examination of the features, understand the causal factors leading to impact and correlation analysis of the sociodemographic characteristics of Local Authorities and furlough levels.

- **Research Objective 2**

  Following the first objective, the second and the most important aim is to predict the impact on business.

  Based on the socio-demographic factors, the objective was to estimate the number of furlough jobs of the Local Authorities of England and to determine the main causal factors predicting the impact using the regression analysis.

3.1 Data Collection

To fill the knowledge gap in previous literatures and studies and to achieve the research objectives, secondary data has been used. This study combines the data of the number of Covid-19 deaths and cases for local areas in England with the data from Index of Multiple Deprivation (IMD) and Cumulative Job Retention Scheme (CJRS) in order to understand their relationships and estimate the impact.

The data of the new confirmed covid-19 cases and deaths by registration date on local authority district level are from Public Health England.

Data of population density are from ONS 2020 population estimates. Data of deprivation are from the 2019 English Index of Multiple Deprivation (IMD) on the Lower Layer Super Output (LSOA) level, which aggregates the following 7 main domains of deprivation: Income, Employment, Education, skills and training, Health, Crime, Barriers to housing and service,

and living environment. Data of Cumulative Job Retention Scheme (CJRS) stating cumulative number of employments on furlough was collected from UK government available open datasets. Data of geographic boundaries of the local districts of England was fetched from the ONS 2019 in the shapefile format. The whole datasets were aggregated on Local Authority District Code to proceed for robust data analysis.

3.2 Data Cleaning

Data cleaning is the process of removing duplicate, missing, corrupted and incorrectly formatted values from the data so that the data analysis can be done in a more robust manner and accurately. To perform the descriptive examination and predictive analysis of features, the data has been thoroughly checked, if it contains any missing, duplicate, negative, negative values or have presence of any outliers (the extreme values). Post data-cleaning, the datasets have been merged so that the new data is ready to predict the impact.

3.3 Data Pre-processing

In order to conduct the regression analysis for estimating the impact and understanding the relationships thereupon, it is vital to split the data into two different class i.e., independent and dependent variable.

3.3.1 Independent Variables

The features that have direct influence on target class and can be moulded in according to business problems. In this study, the independent or input variables taken for analysis are –

Index of Multiple Deprivation (IMD)

The indices of deprivation are a measurement of relative deprivation at a local area level across England. Deprivation encompasses the broad range of individual living conditions.

The seven domains aggregating this index were assumed as sociodemographic factors in order to find the significance on the output variable. The decile of indices in a given local authority has been used as it depicts the deprived 10% of neighbourhood nationally and their ranks range from 1 to 10, where 1 is most deprived 10% of LSOAs. Hence, the seven domains have been taken as the independent variables.

Initially IMD data was highly disaggregated into sub regions at district level, which has been aggregated with the help of evaluating the mean (since the average gives more refined insights)

value of each sub-regions to make it ready for analysis at LSOA level. The dataset contains 317 different local authorities of England.

Number of Confirmed Covid-19 cases

The active number of coronavirus cases are given in such time frame. Since the prime theme of this study was covid, therefore it is vital to consider covid cases for the analysis. The data was initially in the disaggregated form of new cases per day per region, later which has been aggregated (using sum function) to determine the total number of active cases at local authority levels across England. Post aggregation, the collected data was found for 315 different regions. To perform the robust analysis at different local area level, it is necessary that cases are determined proportionately to their population density level. Therefore, cases were determined per 100,000 people.

For the regression analysis and descriptive evidence, covid cases which occurred between March 1st 2020 and June 30th 2021 by Local Layer Super Output Area (LSOA) has been used and clustered in 4 different temporal chunks (Mar 1st 20 - June 30th 20, July 1st 20 – Oct 31st 20, Nov 1st 20 – Feb 28th 21, Mar 1st 21 – June 30th 21) for better period-wise examination of covid cases trend.

Hence the cases per 100,000 people for 4 different periods were the final input variables along with the IMD decile indices.

Number of Confirmed Covid-19 deaths by registration date

These are the number of mortalities due to coronavirus. This is the significant demographic feature as it aligns with the context of research aims and objectives of this study.

The data on the mortalities were similar to the data on covid cases. Hence, the same process has been applied to this data to make it ready for the robust data driven analysis. Therefore, the final dataset for the number of deaths contains 304 local authorities of England and the deaths per 100,000 people for 4 different chunks has been used as the final independent variable alongside data on IMD indices and cases.

3.3.2 Dependent Variable

This is the target or output variable which was tested, measured, and manipulated by the independent variables in order to attain the objective. Hence cumulative number of furlough jobs has been considered as the dependent and estimation feature.

Cumulative number of furlough jobs

The data gathered from Coronavirus Job Retention Scheme (CJRS) by UK govt states that the number of people at different local authority levels has taken temporarily leave from their work or has been furloughed by their employers due to covid-19. This has been used as a proxy data to estimate the business impact at furlough levels. The dataset holds cumulative furlough jobs of 247 regions at LSOA level.

The cumulative furlough jobs have been calculated proportionately to the population size at LSOA level to make the analysis unbiased.

Thus, cumulative furlough jobs per 100, 000 people has been taken as the final output variable for the analysis.

3.4 Descriptive Analytics

Under descriptive examination of the variables, the regions have been distributed on the scale of most to least influence of the pandemic outbreak and featured with the trends, patterns, and their exploration through statistical and correlation analysis.

The daily cases and deaths for 4 periods and their total number for the respective regions have been computed via SQL query using the SQLDF library and aggregation function (sum ()) in order to examine the regional infection of covid-19.

Similarly, using SQLDF library, the most and least deprived areas have been determined based on each domain of Index of Multiple Deprivation decile (income, employment, education, health, crime, housing and living environment) and employment furlough level.

An interactive line chart visualization displays the daily progression rate of cases and deaths in a given time across the 315 local regions of England. If the cycle in line chart is going up, the growth of cases is accelerating and vice versa. Multiple line charts have been employed to visualize the cases and deaths of each chosen 4 temporal chunks.

Bar charts represent information of deprived regions based on seven domains of Index of Multiple Deprivation and the areas infected by the pandemic outbreak on the scale of most to least. Each bar in the chart shows the name of the LSOA regions on x-axis (y-axis in horizontal bar) and deprivation deciles (1-10) and cases/deaths per 100,000 respectively on y-axis (x-axis on horizontal bar). The fadedness of colour on bar chart depicts the influence rate from high to low.

An interactive geospatial map visualization displays the IMD decile and total number of cases and deaths of each region where the highlighted markers represent the most deprived and infected areas according to the respective context. Clicking on the areas of map shows the name of that region and their respective infection rate. The colour scale used in the map depicts the deprivation rate for ex. light colour shows the most deprived or impacted area and vice versa. The map and markers have been visualized with the help of geometry coordinates of local regions using the folium library.

To expand the examination of each feature and determine their relationships and behaviour, the good practice is to perform correlation analysis of each variable (independent and dependent). Correlation analysis is the technique to find the statistical relationship and causal effects between the variables and their association with each other.

Since the data hold both ordinal (IMD Deciles) and continuous values, spearman method of correlation has been used to find the relationship as it evaluates the strength of monotonic relationship between variables and can be used with both ordinal and continuous data. It has been computed using the corr function of pandas library.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Fig 1 – Correlation Formula (Glen, n.d.)

Correlation matrix visualization represents the correlation coefficients range from -1 to 1 between each variable, where the coefficient towards 1 depicts strong relationship, less than 0 shows weak correlation and 0 means no relationship. The light colour in the matrix represents high correlation whereas dark colour shows weak relationship.

Pairplot visualization represents the linear trend and bond between the features and also gives the idea of skewness into the data.

An interactive joint plot visualization reveals the bond between two variables and describe their individual distribution in the data on the same plot. It has been plotted using seaborn library to analyse the relationship more effectively.

Finally, skewness has been checked using skew function in order to find the distribution and degree of asymmetry of the data. Distplot has been used to visualize the skewness.

3.5 Predictive Analysis

After collecting, cleaning, pre-processing, and performing the descriptive examination of features of the data, the next step towards the second goal of the study was to prepare the model to predict the impact of covid-19 on cumulative furlough jobs.

Regression analysis has been employed to build the model, firstly this technique is widely used to determine the causal relationships between input and target variables and secondly, it is used for anticipation and forecasting and indicates the impact.

3.5.1 Building and training model

Various regression models have been implemented in order to determine the working model that fits best with the given dataset. The models tested during the analysis are as follows –

Ridge Regression

It is a technique used to estimate the coefficients of multiple regression algorithms when the presence of multicollinearity is high among the independent variables. The formula below computes the coefficient value of the regression.

$$Y = XB + e$$

Linear Regression

It is the most acceptable and widely used technique for predictive modelling and establishing the relationship between explanatory and response variables. Since the research objective aligns with this algorithm and therefore has been used to test whether it fits with the data. Below is the formula to calculate the linear regression.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

Decision Tree Regressor

Decision Tree is the supervised learning algorithm used to predict the target class by splitting the data into small subsets, in a tree like structure. It is used to predict the continuous quantitative data.

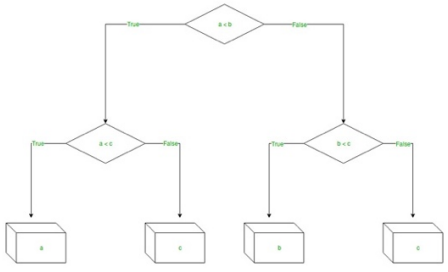


Fig 2 – Decision Tree Structure (GeeksforGeeks, 2018)

Support Vector Regressor

It finds a hyperplane such that loss is minimized. It acknowledges non-linearity of the data and provides a robust prediction model.

K Nearest Neighbors Regression

KNN algorithm is a non-parametric technique used to recognize the patterns and stores available cases and forecast the target value based on the similarity measures (distance function). The knn can be calculated from the formula in the figure no. ().

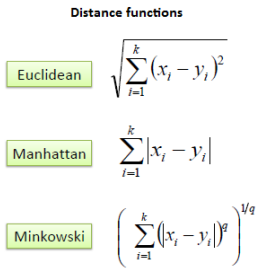


Fig 3 – KNN Regression Formula (www.saedsayad.com, n.d.)

Random Forest Regression

It is an accurate and powerful supervised machine learning model that uses ensemble learning technique (method that aggregates anticipation from various machine learning algorithms to make more robust prediction) of regression.

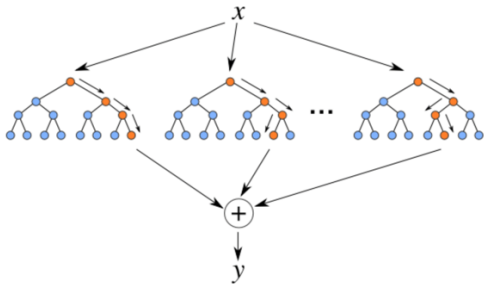


Fig 4 – Random Forest Structure (Bakshi, 2020)

Least Absolute Shrinkage and Selection Operator (LASSO) Regression

It is same as the ridge regression but use the absolute value in the penalty function instead of square value. It is used over regression methods more accurate estimations and can be computed by the formula below:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Fig 5 – Lasso Regression Formula (Stephanie, 2015)

Gradient Boosting Regression (GBR)

It is predictive algorithm used to minimize the loss. It produces a model out of ensemble of weak predictive algorithms. It evaluates the difference (residual) between actual target and prediction value (Paperspace Blog, 2019)

- **Pipeline** - Pipeline technique of machine learning has been employed to fit the models as it prevents the information leakage and would result in better prediction.
- **Cross Validation** – It is resampling procedure used to evaluate and estimate the skill of machine learning algorithm on unseen data. K-fold technique has been employed based on standard parameters in order to build the robust model.
- **Standardization** – It is a scaling technique used to standardize the range of input features where the data values were centered around the mean with a unit standard deviation. StandardScaler has been employed to scale the data.
- **Skewness** – During analysis, it has been found that the data was not normally distributed and was skewed towards left which may results in biased performance. Hence, log function of Numpy library has been used to fix this problem and later after fitting the model it is inversely transformed into its original form (using exponent function) in order to compare the results.

3.5.2 Evaluation Metrics

Performance metrics is used to evaluate the performance and quality of the statistical prediction model. It estimates the degree of fitting of the algorithm into the data.

Various evaluation metrics has been implemented to check model's performances and to select the best thereupon.

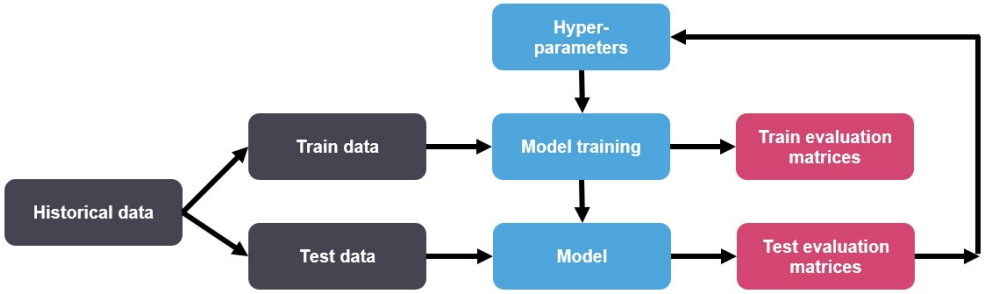


Fig 6 – Evaluation Mechanism (Analytics Vidhya, 2021)

R-squared (R2)

R2 is the proportion of variance in the response variable (outcome) that is explained by the predictor variables. The higher the R-squared value, the better would be the model. The computation of R2 value via formula below -

$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

Fig 7 – R Squared (R2) Formula, (Analytics Vidhya, 2021)

MAE

Mean Absolute Error (MAE) is widely used metric that is more robust to outliers and identifies the absolute difference between the predicted and absolute value of the target variable. The less error represents the better model efficiency. MAE can be calculated by the formula below mentioned in the **figure**

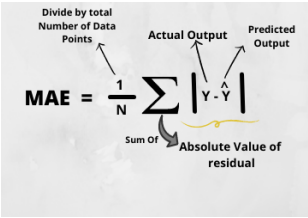


Fig 8 – Mean Absolute Error Formula (Analytics Vidhya, 2021)

MSE

Mean Squared Error or MSE is a well- known error metric for used for regression problems. MSE computes the squared difference between the predicted and actual value.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

Fig 9 – Mean Squared Error Formula (Analytics Vidhya, 2021)

RMSE

Root Mean Squared Error (RMSE) is nothing but the square root of the average of square of all errors or the standard deviation of the prediction errors. It is widely used and robust error metric for numerical predictions.

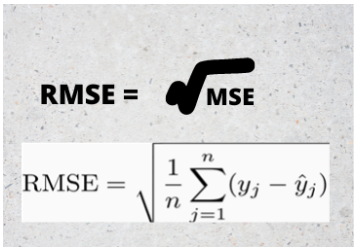


Fig 10 – Root Mean Squared Error Formula (Analytics Vidhya, 2021)

3.5.3 Hyperparameter optimization

Based on the selected performance evaluation metrics, the best model has been chosen for predicting the impact of covid-19 on furlough levels. Hyperparameter tuning is choosing the optimal parameters for the learning algorithm. It is always desired to optimize the hyperparameters as it directs the overfitting and underfitting of the algorithm.

Using grid search library function, hyperparameters has been optimized as it picks out a grid od hyperparameters of the algorithm and measure all of them, it tests out all the hyperparameters and evaluates each of them on loop and selects the best.

Finally, the predicted values have been aggregated to the final data as “Predicted employment furlough” to forecast the regional-wise impact of covid-19.

Scatter plot displays the comparison of actual and predicted employment furlough level and the relationship pattern thereupon. Besides this residual plot has been made as it states how much the line misses the datapoints.

Bar chart represents the comparison of performance of regression models based on evaluation metrics and the most and least affected regions based on prediction of employment furlough. (Figure attached in appendix).

Geospatial map represents the predicted furlough jobs for 247 regions at LSOA level, where the highlighted icons display the prediction of highest furlough jobs in the respective area. Clicking on the areas of map shows the name of that region and their respective predicted furlough level. The colour scale used in the map depicts the furlough rate for ex. light colour shows the high jobs furlough level and vice versa. The black colour depicts the most impacted while green shows the least affected regions.

CHAPTER 4

RESULTS AND DISCUSSIONS

Following the methodology used in the analysis, this chapter elucidates the results of the statistical techniques implemented in the study to achieve the desired research aims and objectives.

4.1 Data Exploration and Pre-processing

The outcome structure of the analysis begins with introspecting the quality of data to determine the false values. While examining the data, no missing, infinite, duplicate or negatives values were detected in each of IMD, Covid-19 cases/deaths, cumulative jobs furlough, population density and geographical boundaries, which determined that the data is in good condition. The reason for its accurateness was the source, as it was fetched from government authorised open datasets where data quality is the prime concern and scrutinized on a regular basis.

COVID-19 patterns

Total cases overtime

The graph below depicts the reported number of coronavirus cases in people in local regions of England from March 2020 to June 2021.

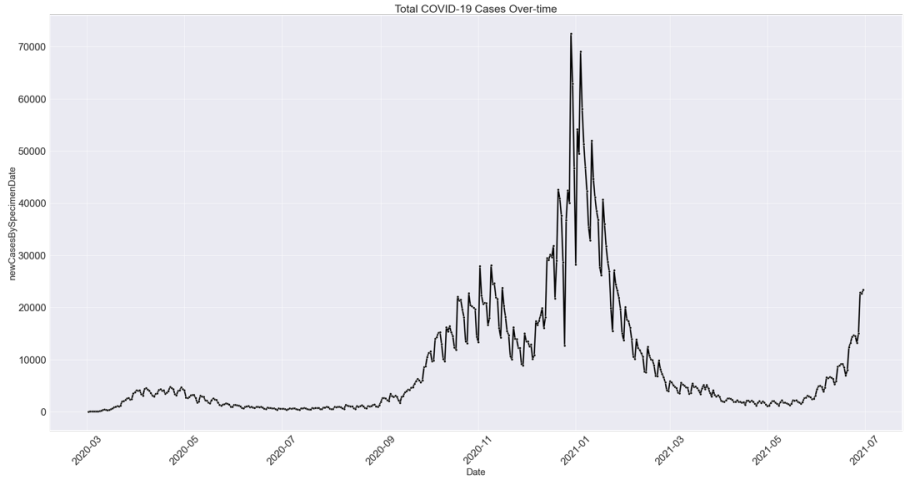


Fig 11 – Total COVID-19 Cases Overtime

The plotted lines represent the number of people who were reported to be in contact of coronavirus in the Local Super Output Layer Area (LSOA) level of England within the given time interval.

Summarising the graph, the number of cases with Covid started increasing in September 2020, where the highest number of people with the infection can be seen in the beginning of 2021.
During the initial stage of the covid, the number of cases increased at slow pace but eventually from the mid of March it rose steadily to around 4500 cases/day until end of May. In the following months, until September, the figure dipped down showing a static trend in the numbers of cases after which it rose significantly reaching a high of 27000 (approx.) till the end of November 2020.
Comparing with the numbers in November, December showed a declining trend in the number of cases, but since the end of 2020, the number of reported cases peaked with around 75000 covid cases per day in January 21. This period was declared as the season of covid where the noted impact of pandemic was large.
Again, in February, there was a plunge in the reported cases' to about 12000. After this point, the cases dipped down again to 3000, but June reported a rising trend in the number of cases till the end of the period.

Total Deaths Over time (March 2020 to June 2021)

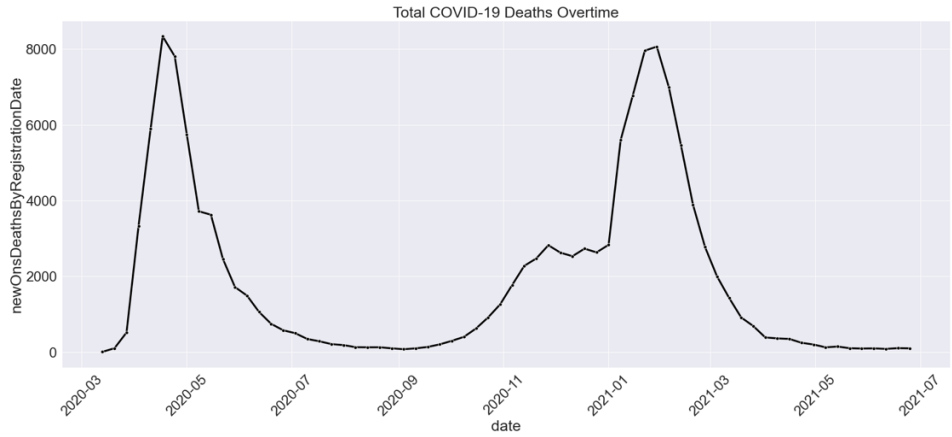


Fig 12 – Total COVID-19 Deaths Overtime

The curve in the figure () depicts the number of mortalities over the time due to the contact of coronavirus in the local regions across England.
Fatalities started increasing in March, where the highest number of people died in the April. The U-shape fatality curve shows the two death waves where significant a number of people died.
Initially, the number of deaths from covid started increasing significantly from the end of March 2020 and reached its peak within 20 days where the reported fatality cases were around 9000/day. Eventually after this peak season the curve shows a considerable downfall in the numbers of deaths for around five months and finally reached to 100 mortalities a day.

However, again after mid-September there was a surge and people started dying due to covid as numbers reaching around 7900 in the January, except the rate fluctuated for about a month from November to end of December. The fatality numbers reported considerable plunge post January and reached to about 80 deaths per day at the end of the period.

Period – 1 (Mar – June 20)

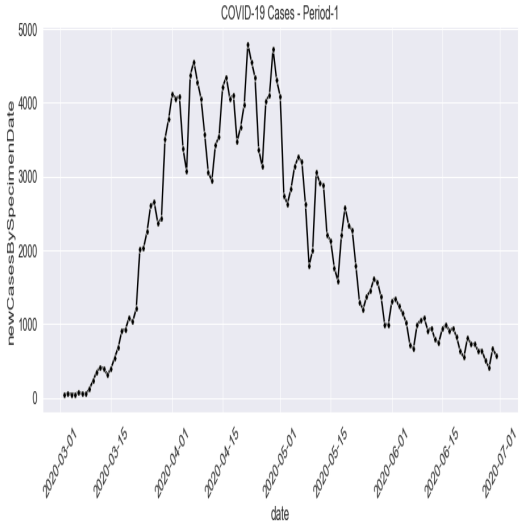


Fig 13a – Total COVID-19 Cases in Period 1

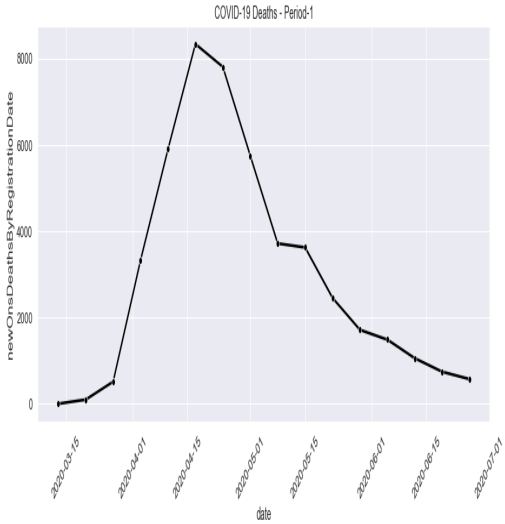


Fig 13b – Total COVID-19 Deaths in Period 1

During the first wave of covid-19, an emerging growth can be seen in the number of cases and fatality during the mid – march, reaching the highest point of 4500 cases and 7900 deaths respectively in the month of April. Post the month of May, the cases reported shows a plunge trend with variations till the end of first period and deaths slowed down after the mid of April.

Period -2 (July – Oct 20)

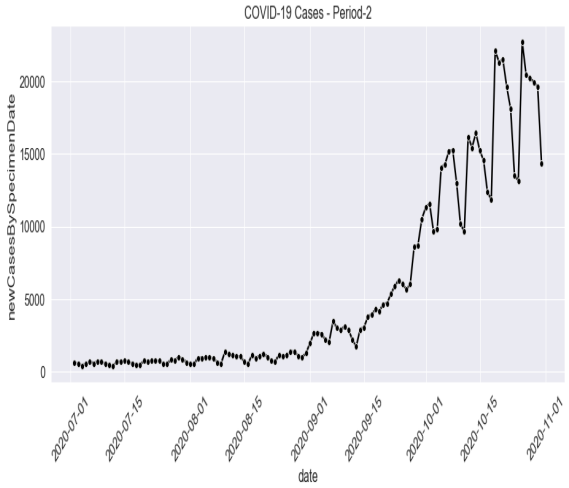


Fig 14a – Total COVID-19 Cases in Period 2

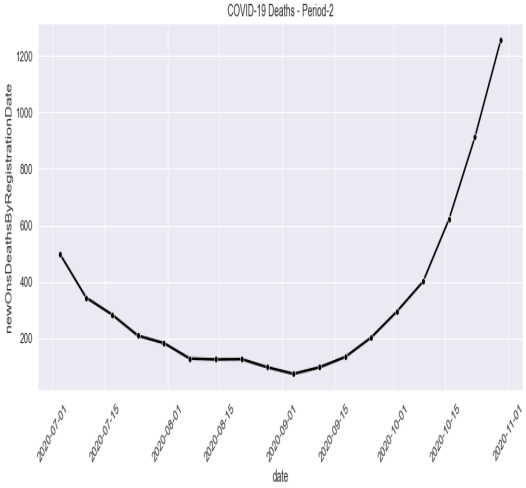


Fig 14b – Total COVID-19 Deaths in Period 2

The cases and death curve depicts a static and declining trend respectively during the first two months of this period, the next time span tells that the infection of coronavirus in the people took a considerable growth for about 25k cases daily and the highest number of deaths noted around 1300 in the month of November. Hence the month of September was an alarm for the authorities and folks to get ready for the impact.

Period 3 (Nov 20 - Feb 21)

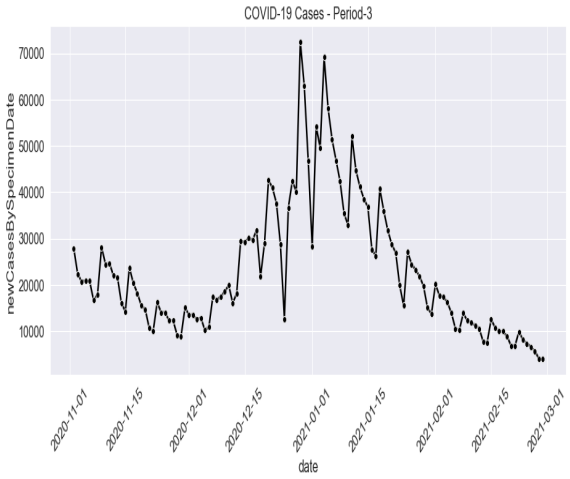


Fig 15a – Total COVID-19 Cases in Period 3

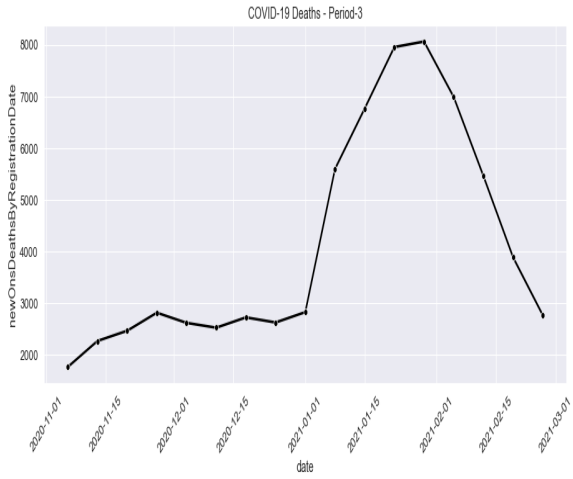


Fig 15b – Total COVID-19 Deaths in Period 3

Next phase of November represents that the cases were plunging and mortalities was low, but over the next few months the height of cases and deaths rose significantly, reported around 75k patients and 7900 deaths in the end of December and in the month of January. In February, the curve declined in the number of cases and fatalities dramatically though a great fluctuation is seen in cases within this fall.

Period -4 (March – June 21)

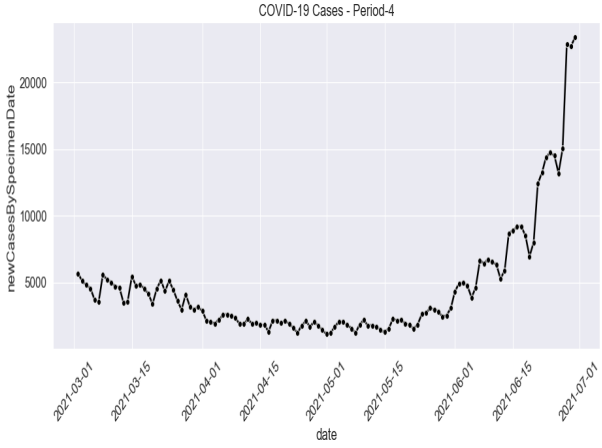


Fig 16a – Total COVID-19 Cases in Period 4

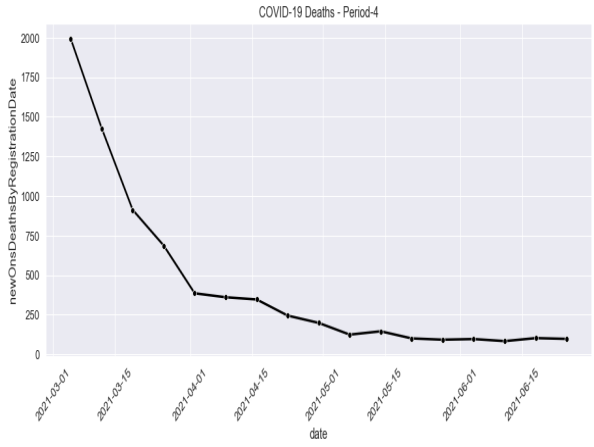
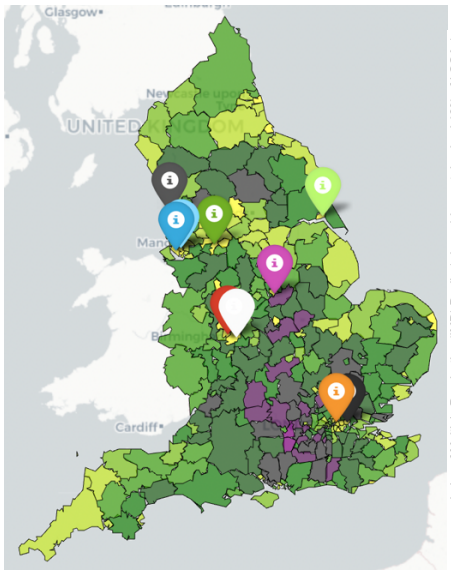


Fig 16b – Total COVID-19 Deaths in Period 4

The last time frame of the overall period represents a different view in the trend of cases and fatalities. A sharp fall can be seen in the reported numbers of patients and deaths except the upward trend in the cases number at the end of the period.

Index of Multiple Deprivation



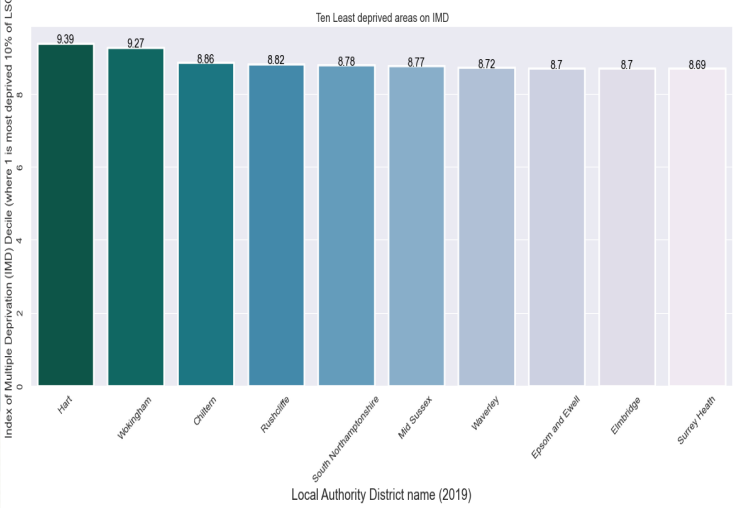


Fig 17a – Most Deprived Areas

Fig 17b – Least Deprived Areas

The map and bar chart represents the most and least deprived regions based on Index of Multiple Deprivation on LSOA level across England.

Overall, figure () shows the order of the areas fall under the rank of IMD deciles on the scale of 1 to 10. The highly noted deprived local areas were Blackpool, Manchester, Knowsley, Barking and Dagenham, Hackney, Liverpool, Sandwell, Birmingham, Kingston upon Hull and Nottingham whereas Hart, Wokingham, Chiltern, Rushcliffe, South Northamptonshire, Mid-Sussex, Waverly, Epsom and Ewell, Elmbridge and Surrey Heath were reported as the least affected area on this index.

The deprived areas which are based on seven distinct domains of IMD are analysed below in order to achieve the aim of thorough examination of data features that explains the sociodemographic characteristics of the local areas, supporting the approach of second research objective of this study.

1- Income –

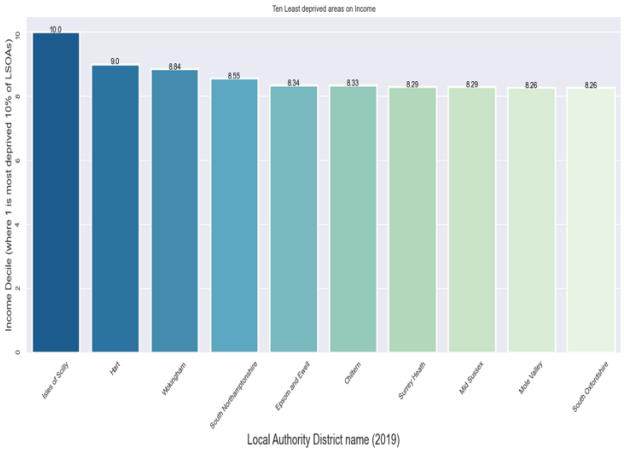


Fig 18a – Least Deprived Areas

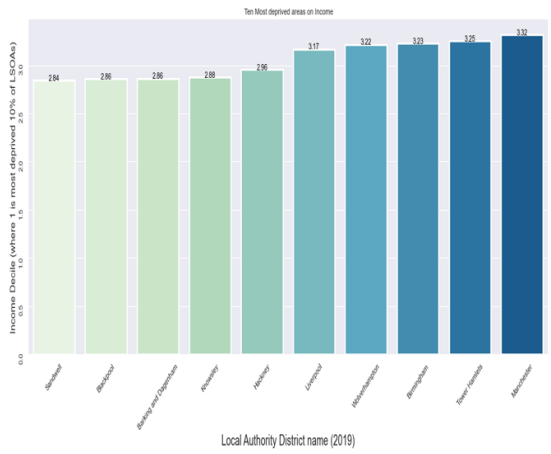


Fig 18b – Most Deprived Areas

The chart describes about the areas whose proportionate population experience the income deprivation. The areas of Isles of Scilly, Hart, Wokingham, South Northamptonshire, Epsom and Ewell, Chiltern, Surrey Heath, Mid Sussex, Mole Valley and South Oxfordshire are the privileged area on the basis of the index while the proportionate people of Sandwell, Blackpool, Barking and Dagenham, Knowsley, Hackney, Liverpool, Wolverhampton, Birmingham, Tower Hamlets and Manchester are highly deprived on income level, where the decile value ranges between 2 and 3.

2- Employment Decile

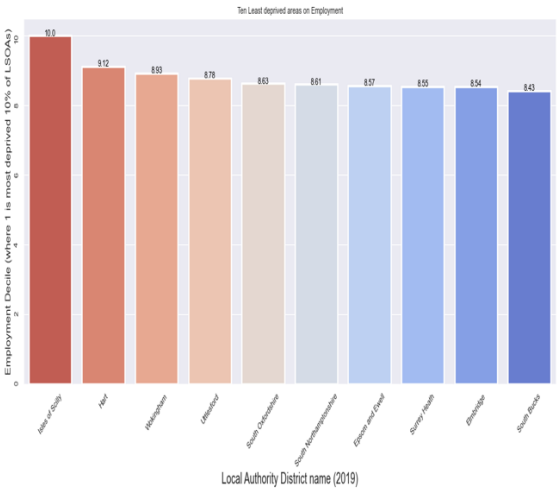


Fig 19a – Least Deprived Areas

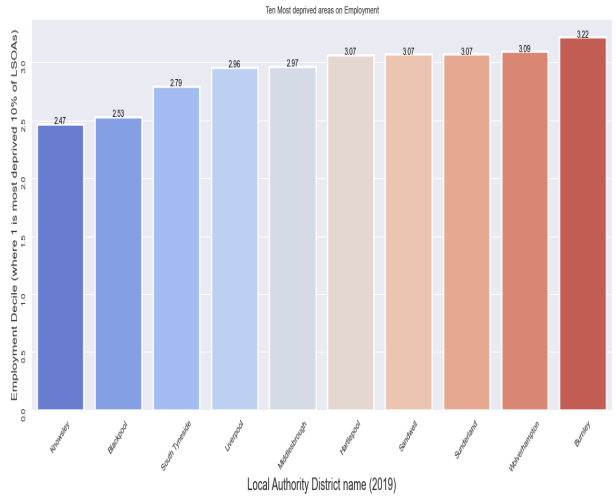


Fig 19b – Most Deprived Areas

The last variable of socio demographic list outlines the deprivation decile of the areas where the population are voluntarily and involuntarily unemployed. Major backward areas of England based on this index covers the Knowsley, Blackpool, South Tyneside, Liverpool, Middlesbrough, Hartlepool, Sandwell, Sunderland, Wolverhampton and Burnley and the

privileged areas where the job opportunities are more and people desires to work are Isles of Scilly, Hart, Wokingham, Uttlesford, South Oxfordshire, South Northamptonshire, Epsom and Ewell, Surrey Heath, Elmbridge, and South Bucks.

3- Education, Skills and Training Deciles

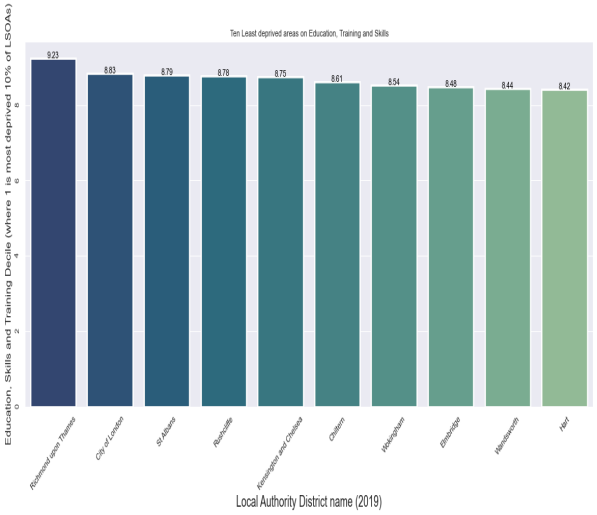


Fig 20a – Least Deprived Areas

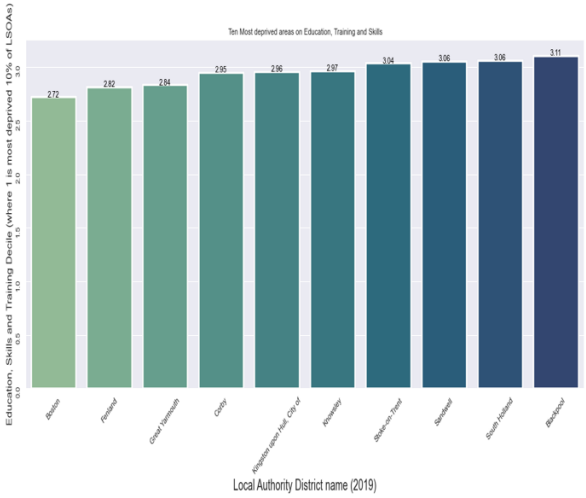


Fig 20b – Most Deprived Areas

The chart in the figure () shows the deprivation deciles in education, skills and training within the local region of England whose population encounters the lack of qualified skills and possess educational disadvantage.

The most and least deprived cities in this socio demographic feature includes Boston, Fenland, Great Yarmouth, Corby, Kingston upon Hull, Knowsley, Stock – on – Trent, Sandwell, South Holland and Blackpool and Richmond upon Thames, City of London, St Albans, Rushcliffe, Kensington and Chelsea, Chiltern, Wokingham, Elmbridge, Wandsworth, and Hart respectively.

4- Health Decile

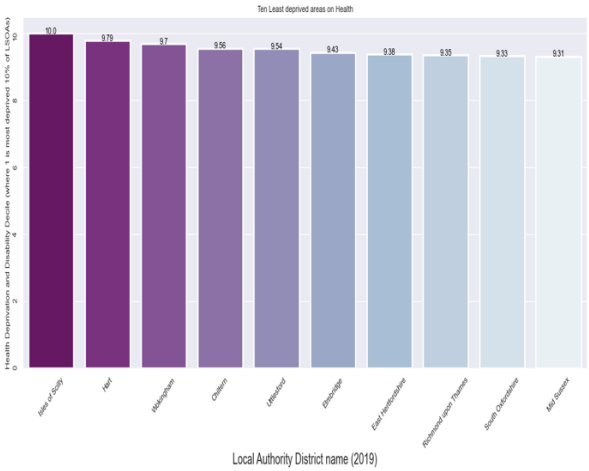


Fig 21a – Least Deprived Areas

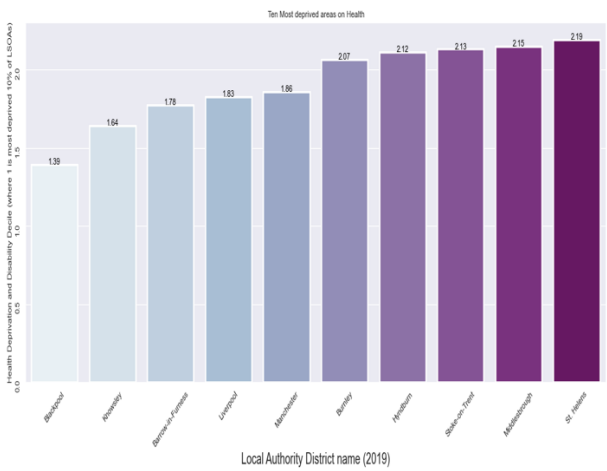


Fig 21b – Most Deprived Areas

The graph of next feature depicts the deciles of quality of health and life of the population based on the local authority level. People of Isles of Scilly, Hart, Wokingham, Chiltern, Uttlesford, Elmbridge, East Hertfordshire, Richmond upon Thames, South Oxfordshire, and Mid Sussex possess good life in term of health whereas the areas of Sandwell, Blackpool, Knowsley, Barrow-in-Furness, Liverpool, Manchester, Burnley, Hyndburn, Stoke-on-Trent, Middlesbrough, and St. Helens have low quality of life as they fall in the upper decile (1-4) of the deprivation index.

5- Crime decile

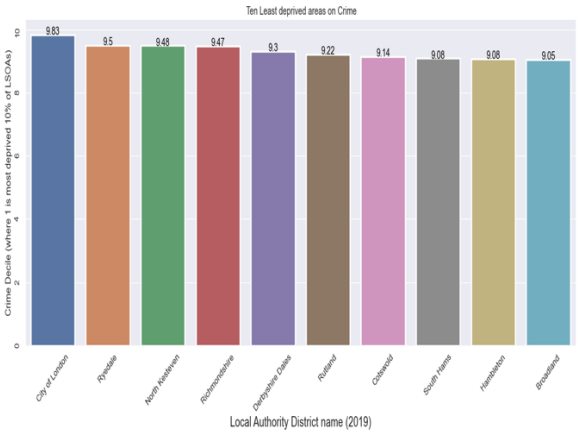


Fig 22a – Least Deprived Areas by crime decile

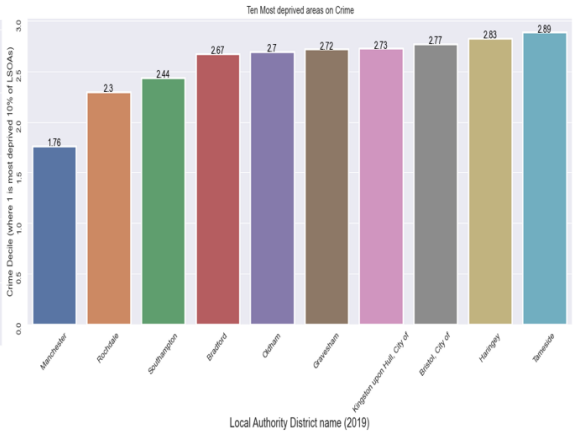


Fig 22b – Most Deprived Areas by crime decile

The Graph of crime decile shows the local areas which records the higher rate of crime in terms of burglary, criminal damage, violence, and theft. London, North Kesteven, Richmondshire, Derbyshire Dales, Rutland, Cotswold, South Hams, Hambleton, and Broadland are the top nine safest cities from crime in England as it falls at bottom of the decile while the nine most deprived regions on this decile are Manchester, Southampton, Bradford, Oldham, Gravesham, Kingston upon Hull, Bristol, Haringey, and Tameside.

6- Barriers to Housing

This chart represents the deprivation decile based on the issues related to the housing accessibility and affordability. The regions that record the most deprived areas covers Newham, Hackney, Barking and Dagenham, Brent, Enfield, Kensington and Chelsea, London, Waltham Forest, Slough, and Haringey while the least impacted areas are Barrow-in-Furness, Hyndburn, Erewash, Pendle, Wirral, Barnsley, Burnley, Sefton, Broxtowe, and Blackpool.

Fig 23a – Least Deprived Areas by housing decile

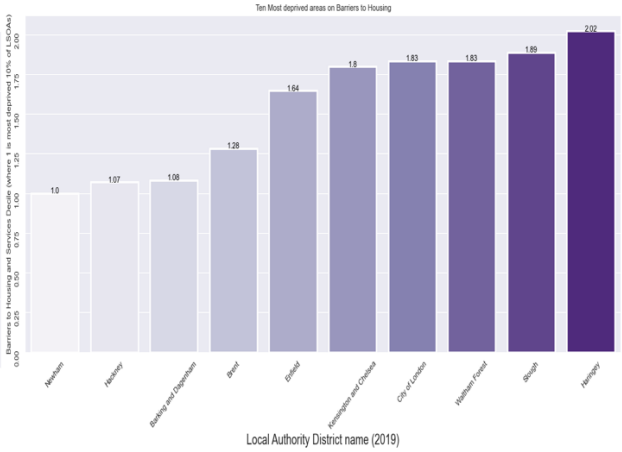


Fig 23b – Most Deprived Areas by housing decile

7- Living Environment

This feature of socio demographic illustrates the deprivation decile of the population of the small areas based on their standard of living. The people of Harlow, Bracknell Forest, Sunderland, South Tyneside, Stockton-on-Tees, County Durham, North Tyneside, Chiltern, Rochford, and Wokingham possess the sophisticated way of living while the residents of Isles of Scilly, Lambeth, Kensington and Chelsea, Westminster, Portsmouth, Birmingham, Pendle, London, Sandwell, and Liverpool have lower Standards of living conditions.

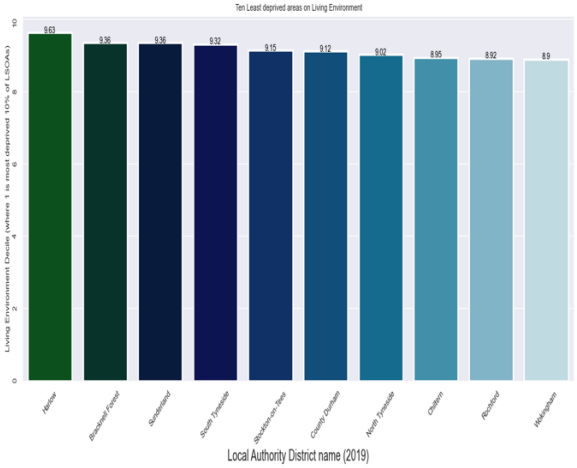


Fig 24a – Least Deprived Areas by Environment decile

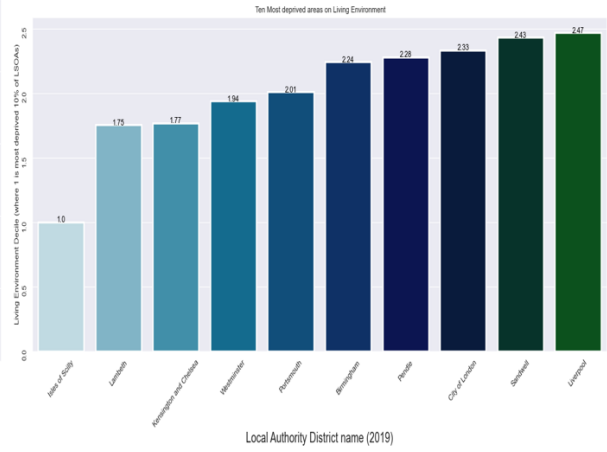


Fig 24b – Most Deprived Areas by environment decile

Total Cases per 100,000

Public Health England reported the COVID-19 new cases and mortality rate by specimen and registration date through June 30, 2021, including London and 315 local areas of England. This data shows a sense of an alarming situation for the population and demonstrates how the deprived areas were hit harder and bears the shock of the pandemic's health impact.

The regions of Blackburn with Darwen, Burnley, Knowsley, Hyndburn, Manchester, Oldham, Middlesbrough, Pendle, Hartlepool, and Salford were highly infected from the virus as highlighted in the figure no. (), however the majority of deaths per 100,000 people were reported in Tendring, Castle Point, Rother, Southend-on-Sea, East Staffordshire, Folkestone and Hythe, Thanet, Burnley, Havering and Eastbourne as mentioned in the figure no. ().

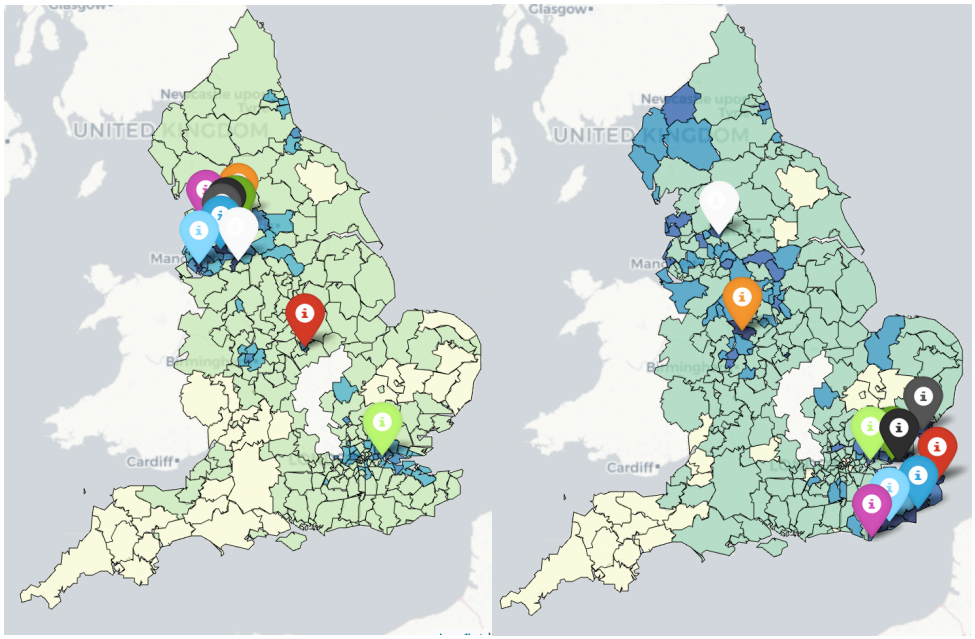


Fig 25a – Affected Regions by Cases *Fig 25b – Affected Regions by Deaths*

Total Cases and Deaths Period-wise

During the first wave of the outbreak, certain areas signalled an alarming situation for its inhabitants. The regions covering Lewisham, Harrow, Sheffield, Brent, King's Lynn and West, Middlesbrough, Norfolk, South Lakeland, Bradford, and Bromsgrove were reported to have the majority of covid infection and fatality rate. In the following period from July to Oct, the upward trend of mortalities and cases has highly affected the big LSOA cities such as Liverpool, Manchester, Leeds, Birmingham, Sheffield, Bradford, and Nottingham.

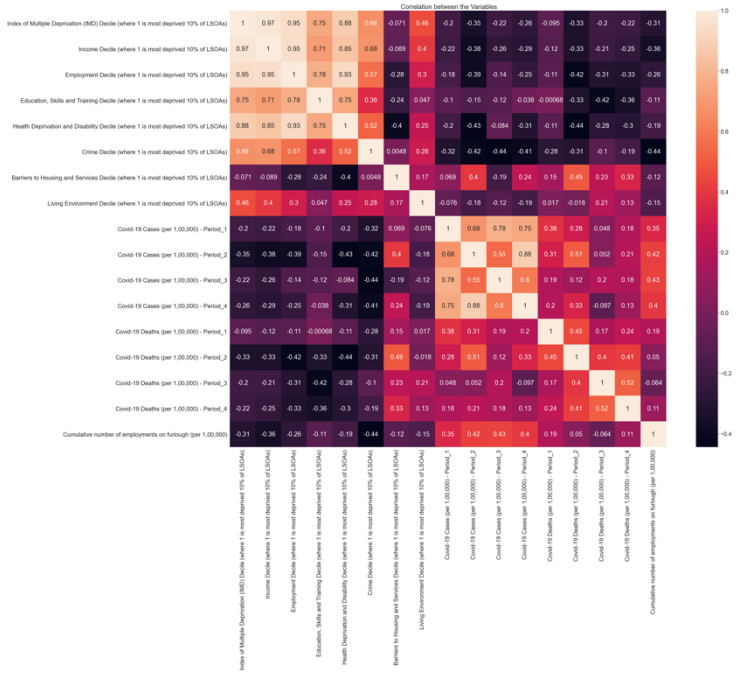
In the third period of coronavirus, which is analysed as the peak wave, the mortalities rate was high in Rother, Castle point, Tendring, Hastings and Eastbourne whereas the regions including Redbridge, Lewisham, Newham, Birmingham, and Sutton shows majority in cases.

The last wave of covid in the given time has a large impact on the people of Manchester, Leeds, Hyndburn, Birmingham, and Bradford while the local regions such as Rochford, Tendring, Dover, Southend-on-Sea and Bury were reported the greatest number of deaths, however it was less than the last two waves. Graphs are attached in the appendix below.

The discussion of the graphs gives the insight of every independent feature and the regions that were majorly impacted and aligns perfectly with the first objective of thorough examination of features.

4.2 Correlation

The correlation matrix represents the relationships between the sociodemographic traits of the local areas of England, Covid-19 Cases and Deaths per 100,000 people and cumulative number of employments on furlough. It can be seen that the coefficients of the deprivation domains of IMD possess strong linear relationship as they are dependent on each other for instance if the unemployment increases (decile towards 1) the income level of an individual plunges (decile towards 1). While sociodemographic features and cases/deaths follows the inverse bond, as the lesser the decile the more would be the impact.



Income

There is a negative correlation (-0.36) between income level of an individuals and the jobs furlough in the local authority. This shows a decline in income as unemployment leads to deprivation, when people have no jobs to do, their earning will automatically reduce.

Employment

The coefficient value -0.26 reveals the weaker relationship between the employment and furlough jobs as the increase in one led to dip in another. The less furlough level depicts the more people on the job.

Education, Skills and Training

Education and Employment furlough reflects the inverse pattern in their bonds with -0.11 as the degree of coefficient. It has been reported that the individuals having skills and educational qualification were less furloughed by the jobs in contrast to the others.

Health

This measure of deprivation is negatively correlated with the furlough levels due to the existence of inverse pattern. The individuals who suffer from sickness and the virus would be furloughed by their jobs more.

Crime

The degree of correlation is far less between crime and furlough levels as they both are the extremities when one doesn't get the work, he commits crime to earn his living.

Barriers to Housing

The relationship of this domain of deprivation holds weak relationship with the employment furlough. The degree of coefficient between them is -0.11, which reflects the inverse pattern.

Living Environment

Living atmosphere is inversely proportional to the level of furlough jobs with the coefficient value of -0.14. The individuals who go on work have healthy standard of living.

Cases – Period-wise

The joint plot represents the relationship between the coronavirus cases and employment furlough. All the cases in the given time frame were positively correlated with the outcome variable, the coefficient values are consistent with the trend of waves of the virus as during the first wave the cases were increasing steadily hence the value is 0.32, eventually after cases rose significantly in the following two periods, so the correlation value i.e., 0.42 & 0.43 respectively, while in the last relief wave of covid, the value shows a dip (0.28) like the period trend.

Death – Period – wise

The analysis of this graph tells a linear relationship between the covid mortalities and employments furlough level. The first two period were positively correlated with the coefficient value of 0.12 and 0.03 respectively while in the third interval it is noted that the correlation value falls negative (-0.04) which reveals the people died would not return to their work. The last period reflected the positive correlation (0.07) between the features.

The aforementioned relationship among the variables reveals the degree of their contribution to predict the impact which is the second objective of the study.

4.3 Data Modelling and Evaluation

After the descriptive and thorough examination of the sociodemographic characteristics of the local regions, the final features that are decided to put into the machine learning algorithm are the IMD deprivation domain, covid cases/deaths per 100,000 of population and the cumulative numbers of employments on furlough. As mentioned in the methodology, based on the evaluation metrics, the performance of the implemented models has been mentioned below –

	MAE	MSE	RMSE	Testing Accuracy
Ridge Regression	0.494630	0.402736	0.634615	0.507869
Linear Regression	0.494994	0.402263	0.634242	0.508446
Decision Tree	0.715545	0.761826	0.872827	0.069070
SVR	0.534862	0.468647	0.684578	0.427327
Knn	0.552345	0.513358	0.716490	0.372691
Lasso	0.587952	0.493198	0.702281	0.397326
Random Forest	0.554428	0.473916	0.688416	0.420888
Gradient Booster	0.525941	0.474045	0.688509	0.420731

Table 1 – Results (Evaluation Matrix)

Mean Absolute Error

In the table (), the mean absolute error showing the difference in the actual and predicted value of target class results the lowest in the ridge and linear regression models i.e., 0.49, whereas the error was the highest for the decision tree regressor which reveals that based on this error model doesn't fit on the data and failed to forecast the furlough jobs accurately. While the error of other regression models such as SVR, Lasso, Random Forest and Gradientboosting were 0.53, 0.59, 0.55 and 0.52 respectively.

Thus, according to this metric ridge and linear regression algorithm anticipated the outcome variable more accurately than other working models.

Mean Squared Error

This error reports the magnitude of risk associated with the predictor. The lesser the error the better would-be anticipation. Based on this performance evaluation technique, the lowest error again would be of linear and ridge estimator with around 0.40, followed by support vector regression model with 0.46 error value. Decision tree algorithm again didn't perform well as with the error coefficient of 0.76.

The estimated figures of MSE value with 0.49, 0.47 reflects the failure of lasso, Random Forest and Gradient Boosting predicting models.

Hence, linear and ridge regression algorithms predicted more precisely as per this evaluation metric.

Root Mean Squared Error

RMSE reflects the quality of anticipation by the applied machine learning models. According to this performance evaluator linear and ridge regression predictors resulted the best output with the same error value of 0.63. Lasso and Decision tree algorithms were evaluated to be the worst with the 0.7 and 0.8 error value respectively. This error represents the distance of prediction from the actual value of the target class. So, the difference was least in the linear and ridge regression model aligning with the errors discussed.

R Squared (R2)

This is the crucial evaluation metric revealing the accuracy of the model in predicting the outcome variable, cumulative furlough jobs in this case. Based on this metric the linear and ridge regression both almost performed the same with the 51% and 50% prediction magnitude of the outcome variable, whereas the decision tree and lasso model doesn't seem perfect for this dataset and noted the bad performance with only 0.0.6 and 0.39 be the r2 value.

The algorithms including random forest and gradient boosting were nearby same with the prediction accuracy of 0.42.

Based on the error values and quality of prediction, the analysis of the applied models suggests that ridge and linear regression algorithms were more accurate in predicting the business impact of covid-19 than other machine learning models. Hence, the linear regressor has been chosen for the final forecast of the effect as it is the best model for prediction and correctly anticipates more than half of furlough class.

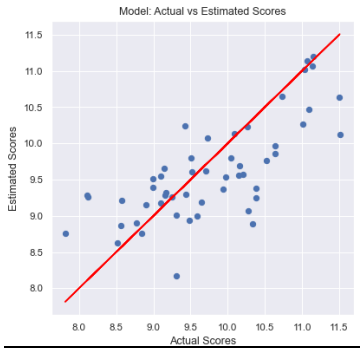


Fig 26 – Actual Vs Predicted Scatter Plot

The figure elucidates the data points of the prediction model made where dots represent the anticipated datapoint.

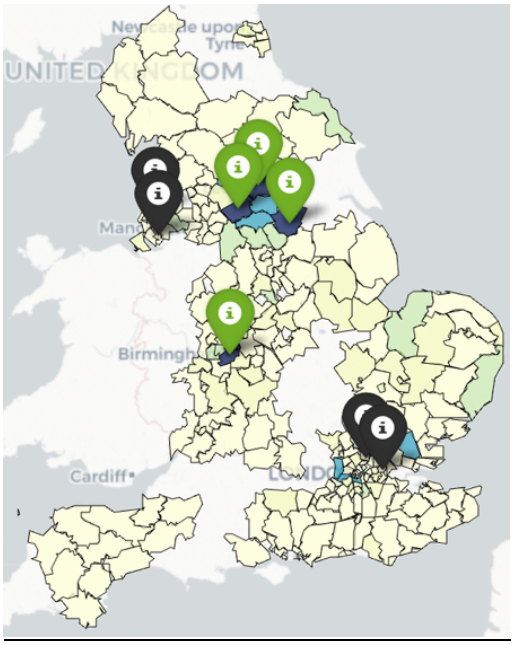


Fig 27 – Most and Least Impacted Regions by furlough jobs

The results suggests that the businesses of local areas of England that will be majorly impacted based on the difference between the actual and predicted level of number of employments on furlough includes Barking and Dagenham, Enfield, Knowsley and West Lancashire whereas the least impacted areas are Leeds, Birmingham, Kirklees and Doncaster.

Thus, predicting the furlough impact and determining the highly impacted regions completes are finding of second objective of this dissertation.

CHAPTER 5

CONCLUSION, RECOMMENDATION AND LIMITATION

The study details the cumulative furlough of jobs in England per 100,000 of population from March 1st, 2020, to June 30th, 2021. The descriptive examination performed on the sociodemographic characteristics of local authority regions results in predicting the impact of Covid-19 outbreak on businesses which shows that a part of population having bad living conditions and worse level of income and lifestyle reported more coronavirus infection and mortalities per 100,000 people. For employment furlough, barriers to housing is less correlated but there is a clear correlation with education, skills, crime, with the cases reported in the period 2 and 3.

Using the developed machine learning model for predicting the impact, the outcomes of regression analysis reports that the large proportion of businesses located in the suburb of London and north-west of England in contrast to Yorkshire were more vulnerable to employment furlough.

A massive variation has been identified in the predicted employment furlough level, which expresses the difference of human resource allocation.

Recommendation

The study has been undertaken with the motivation to suggest the local authorities to overcome the business impact from the pandemic outbreak.

It has been exhibited that by using local and regional data, the predictive model can be utilized to direct the local medical demands and potential, strategy making and public health decisions to mitigate the impact of covid 19 on the local businesses. The upcoming future waves could possibly affect the business of the nation, this will help to timely ensure the organization and commissioning of services. The algorithm can be used by the government for planning the potential schemes and strategies which will help tackling the business impact.

At few moments government should declare monetary schemes uniformly to all the regions where the businesses of less vulnerable areas suffer, which will indirectly affects the economy of the nation. Using this predictive model, the government can offer schemes in a proportionate manner by categorising the areas based on their depravity, which will not only aids the government to allocate their resources optimally but also empower the businesses to sustain.

Limitations

The study on impact of covid-19 is backed by following drawbacks –

- In the development of the predictive model for thorough data driven analysis, it is crucial to have enough data sample size, which allows the algorithms to better examine the trends and pattern from the data. The sample size was very small to analyse the impact, if it could have been better the anticipated regional-wise employment would have resulted more accurately then the noted results.
- To determine the real impact, it is better to have more realistic data to explain more clearly. In this study, the data on cumulative furlough of jobs was used as the proxy of business impact, which predicted only one side of the business not business as a whole.

REFERENCES

Achim, M.V., Safta, I.L., Văidean, V.L., Mureșan, G.M. and Borlea, N.S. (2021). The impact of covid-19 on financial management: evidence from Romania. *Economic Research-Ekonomska Istraživanja*, pp.1–26.

An, C., Lim, H., Kim, D.-W., Chang, J.H., Choi, Y.J. and Kim, S.W. (2020). Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. *Scientific Reports*, [online] 10(1), p.18716. Available at: https://www.nature.com/articles/s41598-020-75767-2.

Analytics Vidhya. (2021). Evaluation Metrics for Your Regression Model. [online] Available at: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

Analytics Vidhya. (2021). Hyperparameter Tuning | Evaluate ML Models with Hyperparameter Tuning. [online] Available at: https://www.analyticsvidhya.com/blog/2021/04/evaluating-machine-learning-models-hyperparameter-tuning/.

Aum, S., Lee, S.Y. (Tim) and Shin, Y. (2020). *COVID-19 Doesn't Need Lockdowns to Destroy Jobs: The Effect of Local Outbreaks in Korea*. [online] RePEc - Econpapers. Available at: https://econpapers.repec.org/paper/nbrnberwo/27264.htm

Bakshi, C. (2020). Random Forest Regression. [online] Medium. Available at: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Bartik, A.W., Bertrand, M., Cullen, Z., Glaeser, E.L., Luca, M. and Stanton, C. (2020). The Impact of COVID-19 on Small Business Outcomes and Expectations. *Proceedings of the National Academy of Sciences*, [online] 117(30). Available at: https://www.pnas.org/content/117/30/17656.

Bongaerts, D., Mazzola, F. and Wagner, W. (2021). Closed for business: The mortality impact of business closures during the Covid-19 pandemic. *PLOS ONE*, 16(5), p.e0251373.

Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of The Total Environment*, p.138817.

CNN, L.A. and A.C. (2020). *Nearly 80% of hotel rooms in the US are empty, according to new data*. [online] CNN. Available at: https://edition.cnn.com/2020/04/08/us/hotel-rooms-industry-coronavirus-trnd/index.html.

Coibion, O., Gorodnichenko, Y. and Weber, M. (2020). Labor Markets During the COVID-19 Crisis: A Preliminary View. *SSRN Electronic Journal*.
Cucinotta, D. and Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Bio-Medica: Atenei Parmensis*, [online] 91(1), pp.157–160. Available at: https://www.ncbi.nlm.nih.gov/pubmed/32191675.

Donthu, N. and Gustafsson, A. (2020). Effects of COVID-19 on business and research. *Journal of Business Research*, [online] 117(1), pp.284–289. Available at: .

Ehlert, A. (2021). The socio-economic determinants of COVID-19: A spatial analysis of German county level data. *Socio-Economic Planning Sciences*, p.101083.

Fairlie, R. (2020). The impact of COVID-19 on small business owners: Evidence from the first 3 months after widespread social-distancing restrictions. *Journal of Economics & Management Strategy*, 29(4).

GeeksforGeeks. (2018). Python | Decision Tree Regression using sklearn. [online] Available at: https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

Ghosal, S., Sengupta, S., Majumder, M. and Sinha, B. (2020). Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4), pp.311–315.

Glen, S. (n.d.). Spearman Rank Correlation (Spearman's Rho): Definition and How to Calculate it. [online] Statistics How To. Available at: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/spearman-rank-correlation-definition-calculate/.

Gregurec, I., Tomičić Furjan, M. and Tomičić-Pupek, K. (2021). The Impact of COVID-19 on Sustainable Business Models in SMEs. Sustainability, 13(3), p.1098

Gupta, S., Montenovo, L., Nguyen, T.D., Rojas, F.L., Schmutte, I.M., Simon, K.I., Weinberg, B.A. and Wing, C. (2020). *Effects of Social Distancing Policy on Labor Market Outcomes*. [online] ideas.repec.org. Available at: https://ideas.repec.org/p/nbr/nberwo/27280.html

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M. and Xiao, Y. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, [online] 395(10223), pp.497–506. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext.

International Labour Organization (2020). *The impact of the COVID-19 pandemic on jobs and incomes in G20 economies*. [online]. Available at: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---cabinet/documents/publication/wcms_756331.pdf.

Lee, C.-C. and Chen, M.-P. (2020). The impact of COVID-19 on the travel and leisure industry returns: Some international evidence. *Tourism Economics*, p.135481662097198.

Li, Q., Feng, W. and Quan, Y.-H. (2020). Trend and forecasting of the COVID-19 outbreak in China. *Journal of Infection*, [online] 80(4), pp.469–496. Available at: https://www.journalofinfection.com/article/S0163-4453(20)30095-5/fulltext.

Lozano Rojas, F., Jiang, X., Montenovo, L., Simon, K., Weinberg, B. and Wing, C. (2020). *Is the Cure Worse than the Problem Itself? Immediate Labor Market Effects of COVID-19 Case Rates and School Closures in the U.S*. [online] RePEc - Econpapers. Available at: https://econpapers.repec.org/paper/nbrnberwo/27127.htm

McKee, M. and Stuckler, D. (2020). If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. *Nature Medicine*, [online] 26, pp.1–3. Available at: https://www.nature.com/articles/s41591-020-0863-y.

Mohamed, A.H. (2021). A Literature Review: The Impact of COVID-19 Pandemic on Somaliland Economy. *Open Journal of Social Sciences*, 09(02), pp.54–64.

Mojjada, R.K., Yadav, A., Prabhu, A.V. and Natarajan, Y. (2020). Machine Learning Models for covid-19 future forecasting. *Materials Today: Proceedings*.

Ngepah, N. (2021). Socio-economic determinants of global COVID-19 mortalities: policy lessons for current and future pandemics. *Health Policy and Planning*.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. and Agha, R. (2020). The Socio-Economic Implications of the Coronavirus and COVID-19 Pandemic: A Review. *International Journal of Surgery*, [online] 78(78). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162753/.

OECD, D.O. and (2020). *The impact of the COVID-19 pandemic on jobs and incomes in G20 economies*. [online] www.ilo.org. Available at: https://www.ilo.org/global/about-the-ilo/how-the-ilo-works/multilateral-system/g20/reports/WCMS_756331/lang--en/index.htm

Paperspace Blog. (2019). Implementing Gradient Boosting Regression in Python. [online] Available at: https://blog.paperspace.com/implementing-gradient-boosting-regression-python/.

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B. and Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General Psychiatry*, 33(2), p.e100213.

Roda, W.C., Varughese, M.B., Han, D. and Li, M.Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic? Infectious Disease Modelling, [online] 5, pp.271–281. Available at: https://www.sciencedirect.com/science/article/pii/S2468042720300075?via%3Dihub.

Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.-W., Aslam, W. and Choi, G.S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access*, [online] pp.101489–101499. Available at: https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-680090.

Sá, F. (2020). *Socioeconomic Determinants of COVID-19 Infections and Mortality: Evidence from England and Wales*. [online] . Available at: http://ftp.iza.org/pp159.pdf.

Saba, A.I. and Elsheikh, A.H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Safety and Environmental Protection*, 141, pp.1–8.

Satu, M.S., Howlader, K.C., Mahmud, M., Kaiser, M.S., Shariful Islam, S.M., Quinn, J.M.W., Alyami, S.A. and Moni, M.A. (2021). Short-Term Prediction of COVID-19 Cases Using Machine Learning Models. *Applied Sciences*, [online] 11(9), p.4266. Available at: https://www.mdpi.com/2076-3417/11/9/4266.

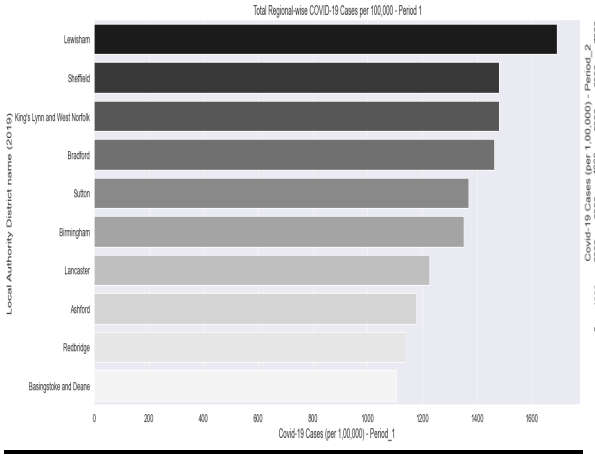
Stephanie (2015). *Lasso Regression: Simple Definition*. [online] Statistics How To. Available at: https://www.statisticshowto.com/lasso-regression/.

Sujath, R., Chatterjee, J.M. and Hassanien, A.E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*.

Wang, L., Li, J., Guo, S., Xie, N., Yao, L., Cao, Y., Day, S.W., Howard, S.C., Graff, J.C., Gu, T., Ji, J., Gu, W. and Sun, D. (2020). Real-time estimation and prediction of mortality caused by COVID-19 with patient information-based algorithm. *Science of The Total Environment*, [online] 727, p.138394. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0048969720319070.

www.saedsayad.com. (n.d.). KNN Regression. [online] Available at: https://www.saedsayad.com/k_nearest_neighbors_reg.htm.

Appendix



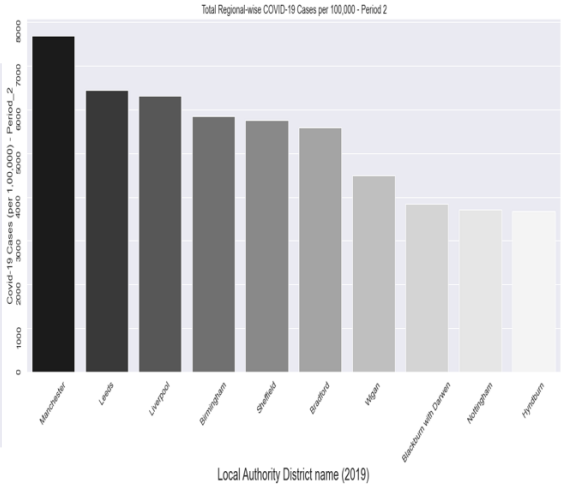
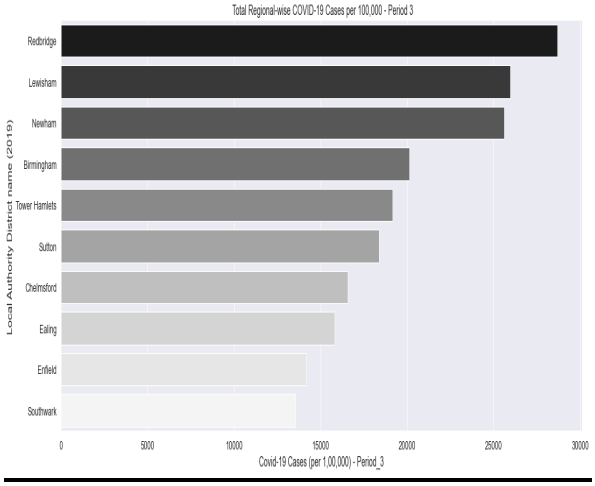


Fig 28a – Least Impacted Regions by cases (Period 1)

Fig 29 – Most Impacted Regions by cases (Period 2)



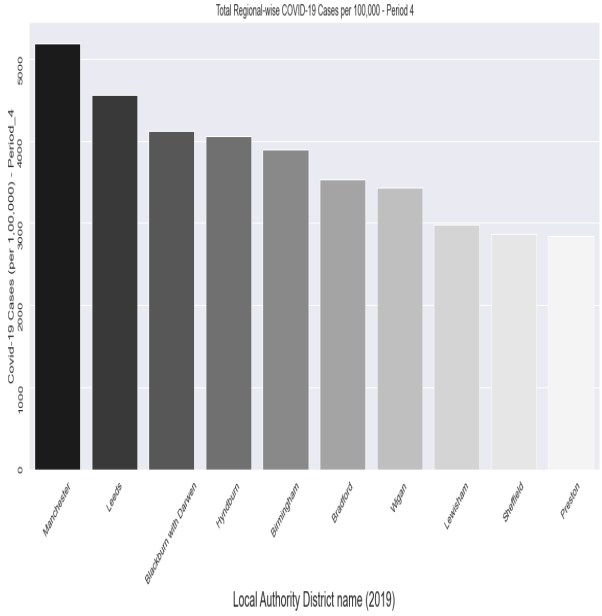
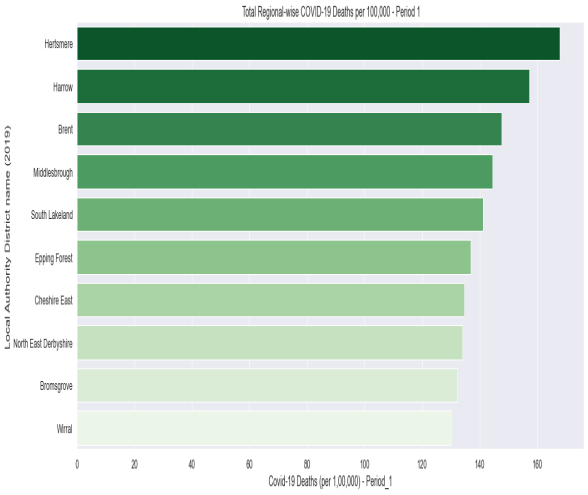


Fig 30 – Least Impacted Regions by cases (Period 3)

Fig 31 – Most Impacted Regions by cases (Period 4)



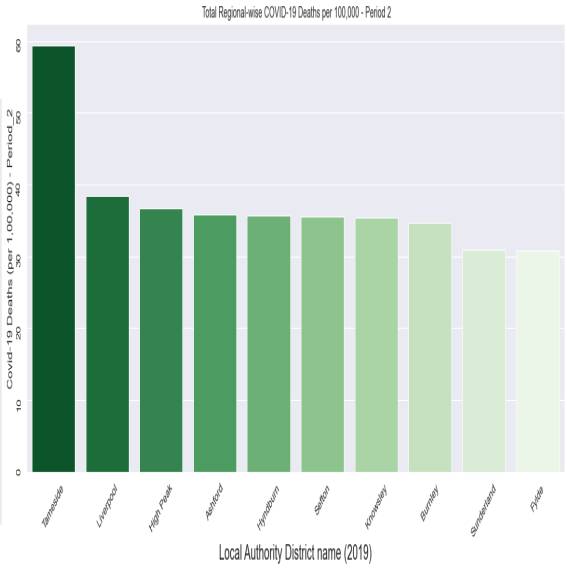
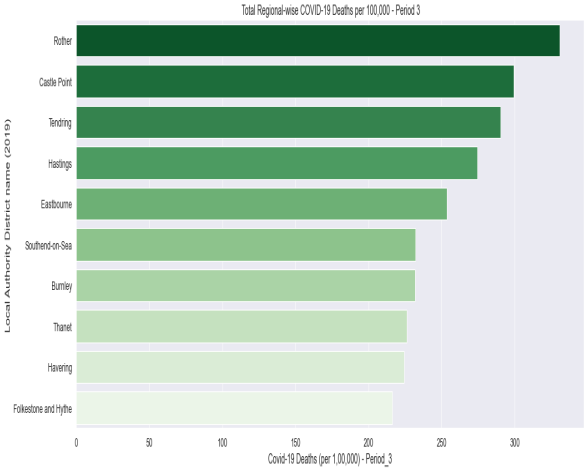


Fig 32 – Least Impacted Regions by deaths (Period 1)

Fig 33 – Most Impacted Regions by deaths (Period 2)



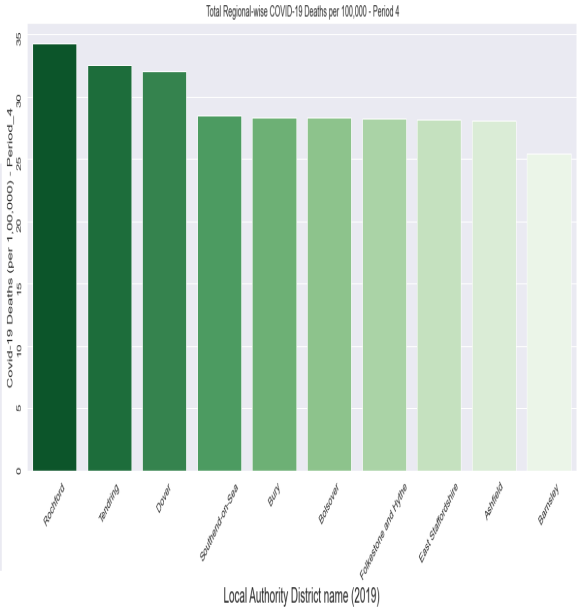
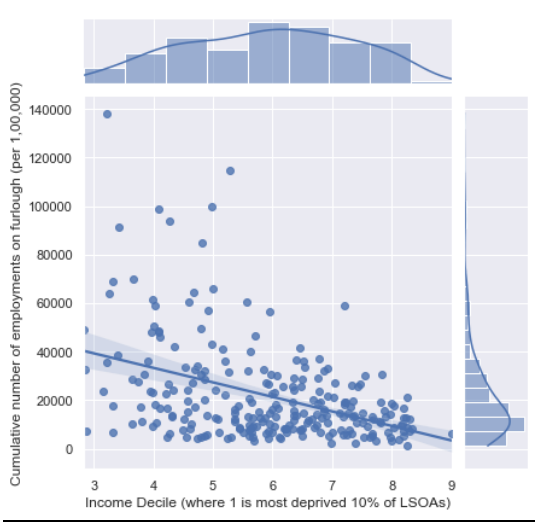


Fig 34 – Least Impacted Regions by deaths (Period 3)

Fig 35 – Most Impacted Regions by deaths (Period 4)



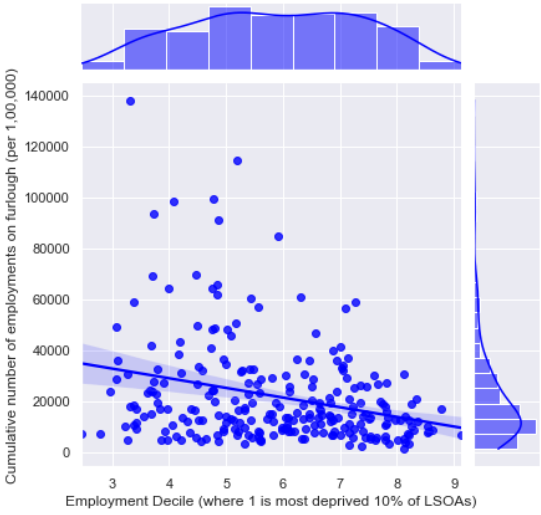
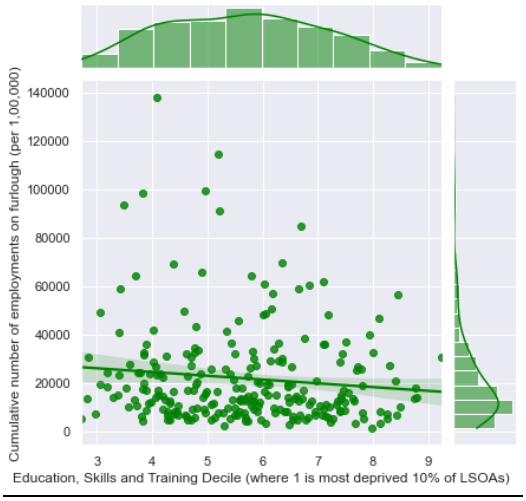


Fig 36 – Correlation Jointplot (Income Decile and Furlough Jobs) *Fig 37 – Correlation Jointplot (Employment Decile and Furlough Jobs)*



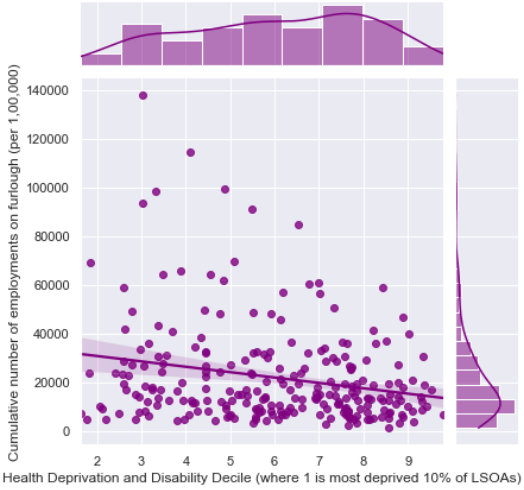
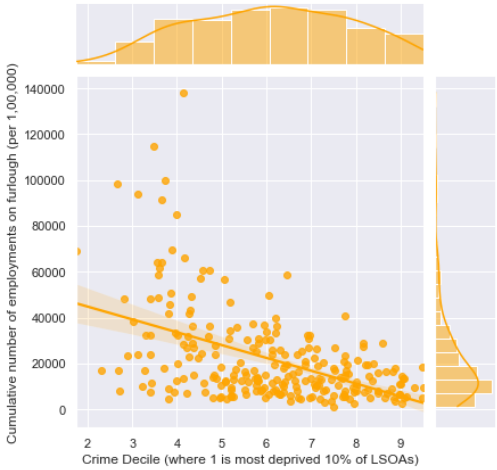


Fig 38 – Correlation Jointplot (Education Decile and Furlough Jobs) *Fig 39 – Correlation Jointplot (Health Decile and Furlough Jobs)*



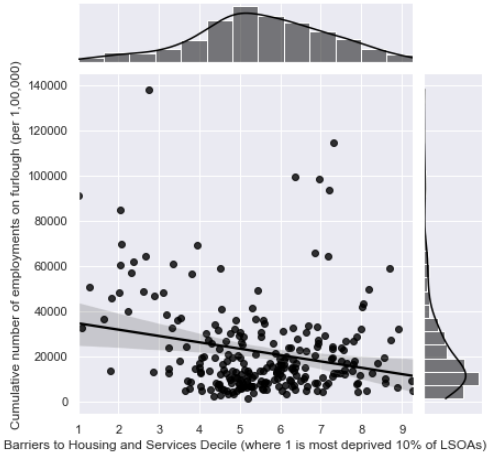


Fig 40 – Correlation Jointplot (Crime Decile and Furlough Jobs) *Fig 41 – Correlation Jointplot (Housing and Furlough Jobs)*

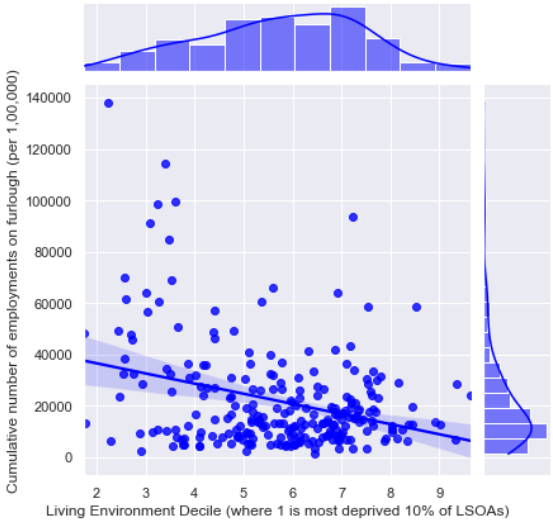


Fig 42– Correlation Jointplot (Environment Decile and Furlough Jobs)

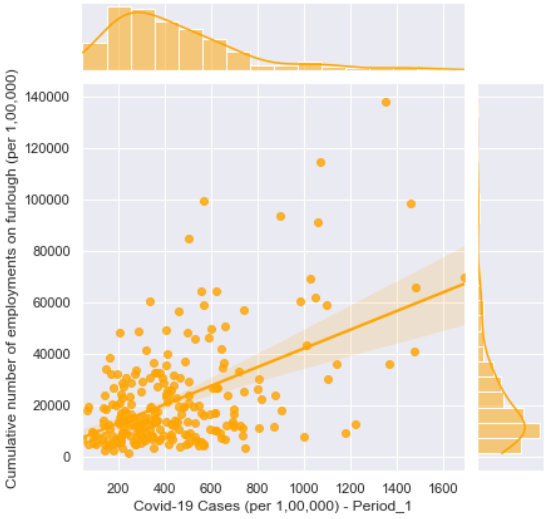


Fig 43 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

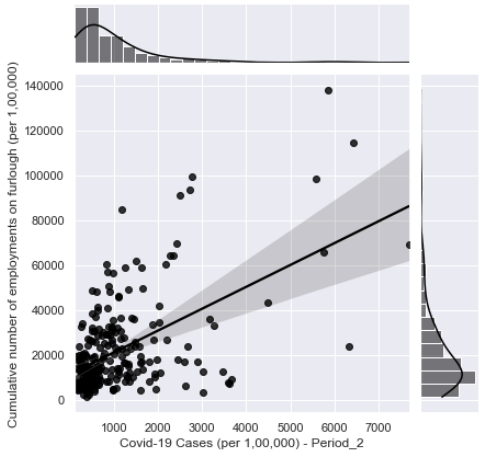


Fig 44– Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

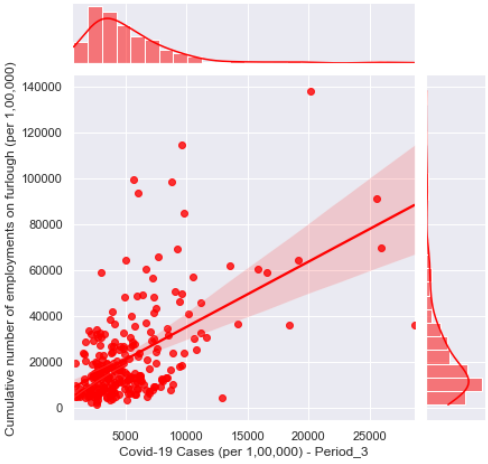


Fig 45 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

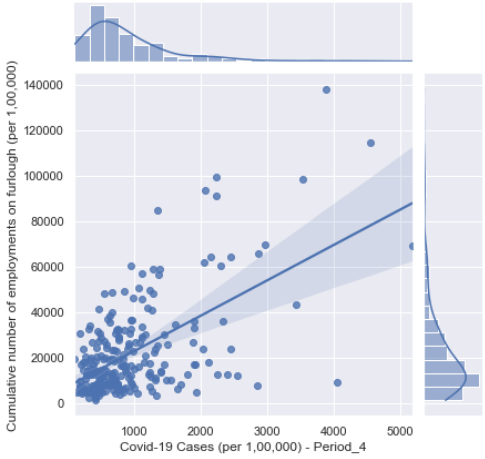


Fig 46 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

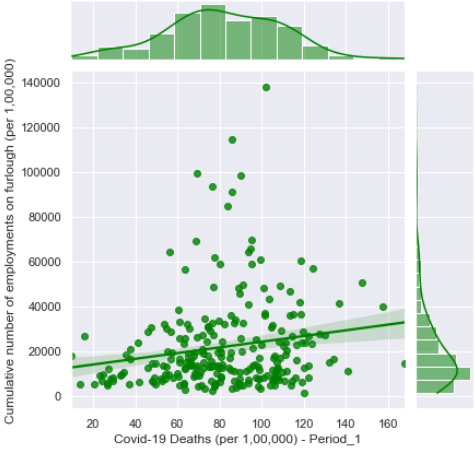


Fig 47 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

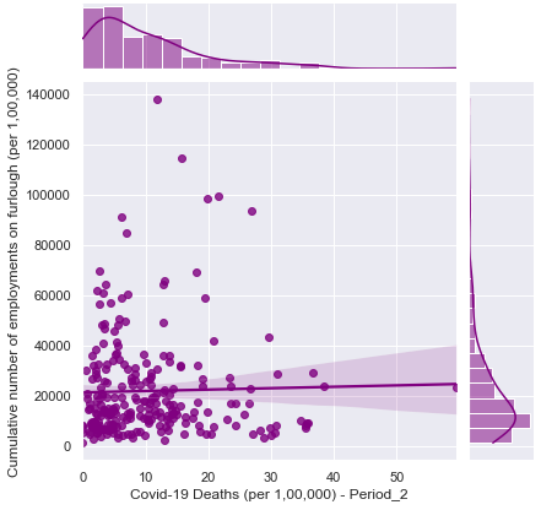


Fig 48 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

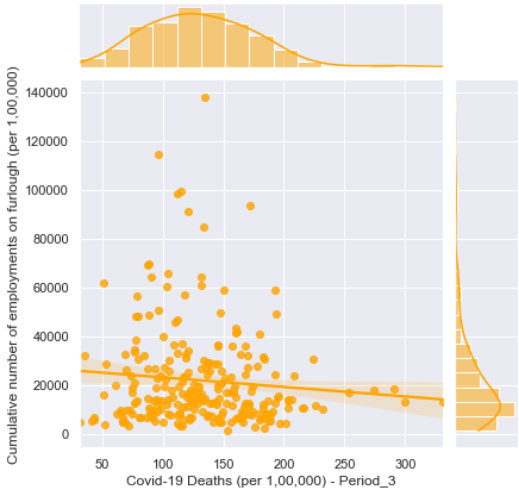


Fig 49 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

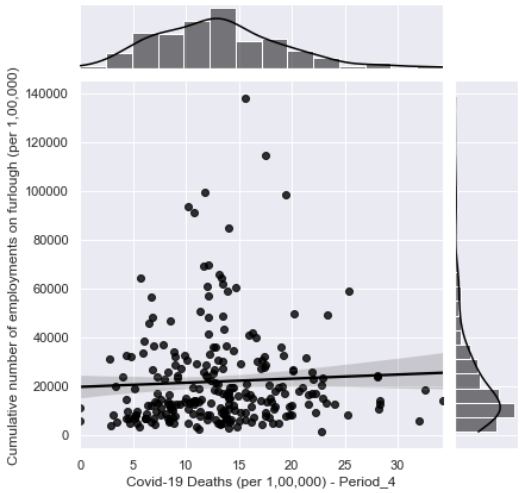


Fig 50 – Correlation Jointplot (Covid Cases Period-1 and Furlough Jobs)

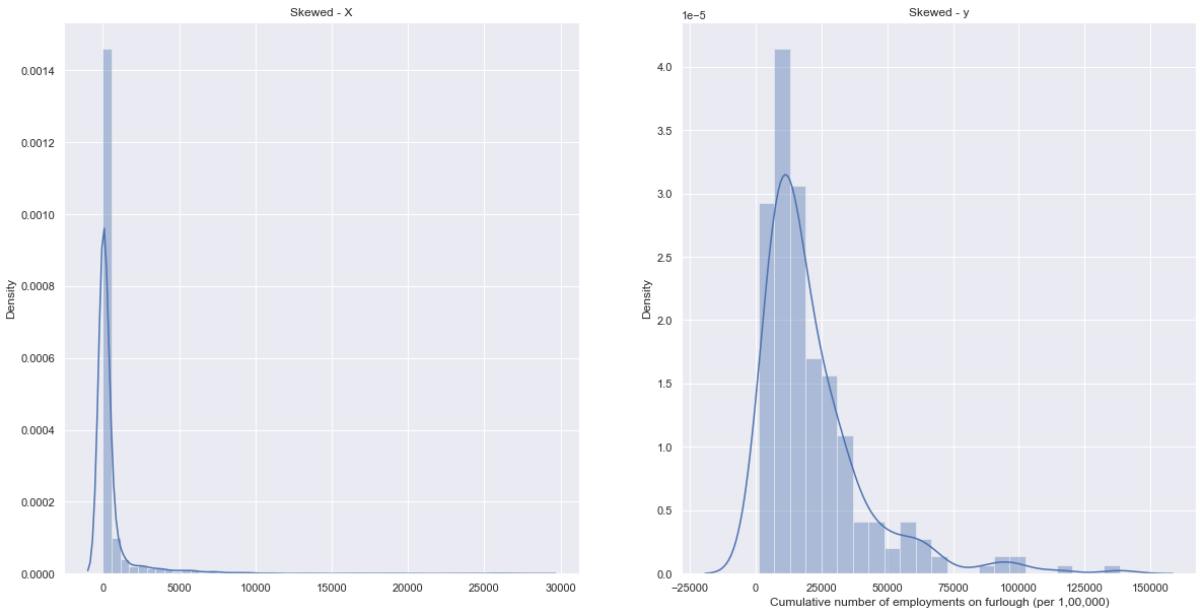


Fig 51 – Skewed Data Distplot

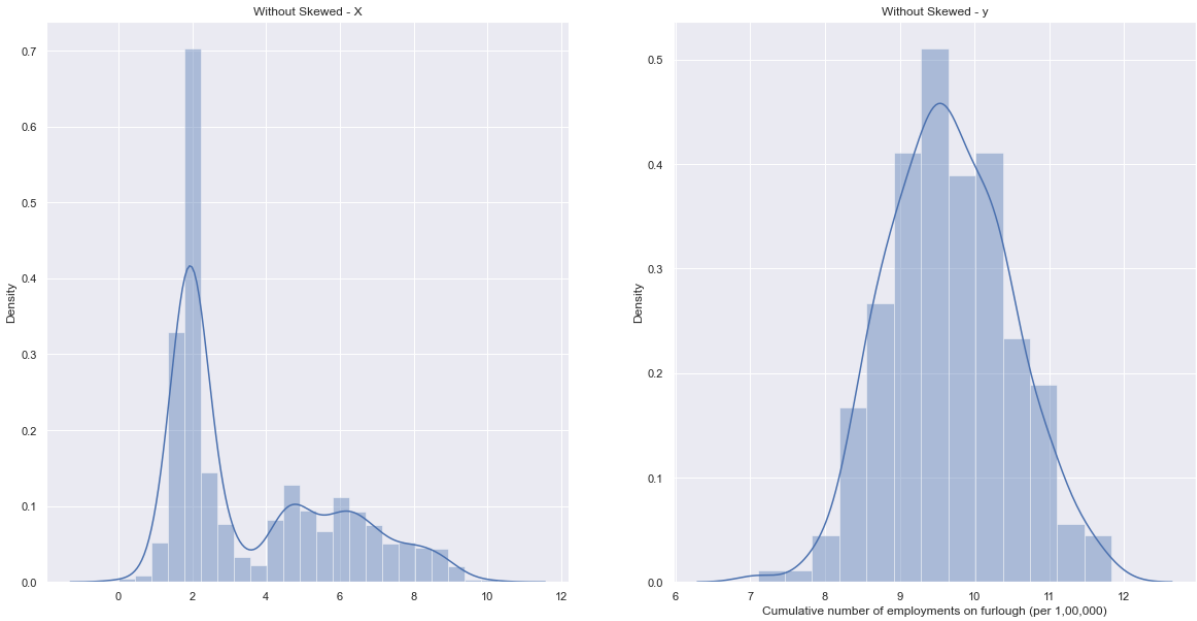


Fig 52 – Unskewed Data Distplot

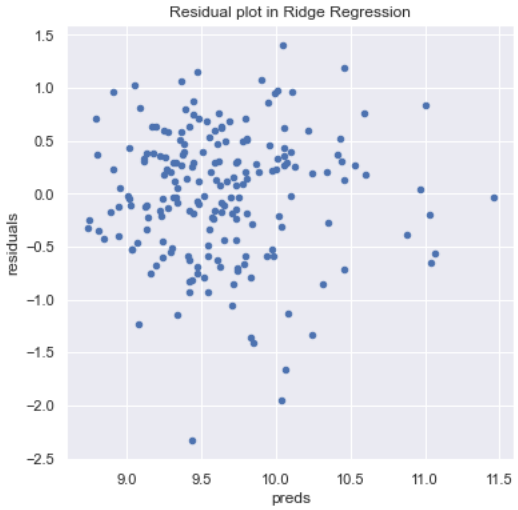


Fig 53 –Residual Jointplot (Ridge Regression)

Fig 54 –Residual Jointplot (Linear Regression)

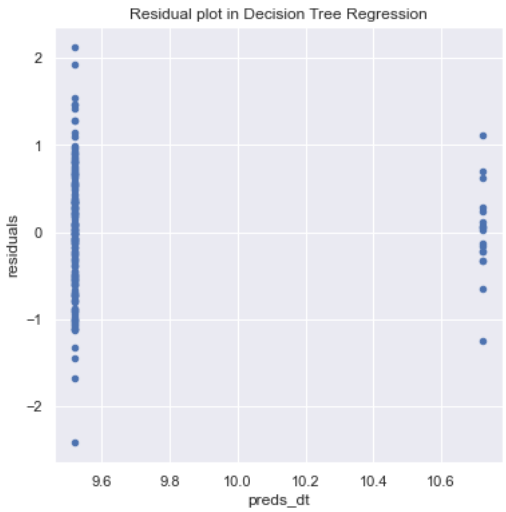


Fig 55–Residual Jointplot (Decision tree Regression)

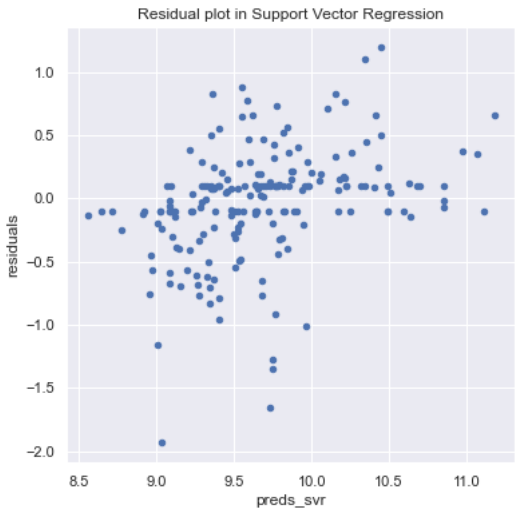


Fig 56 –Residual Jointplot (Support Vector Regression)

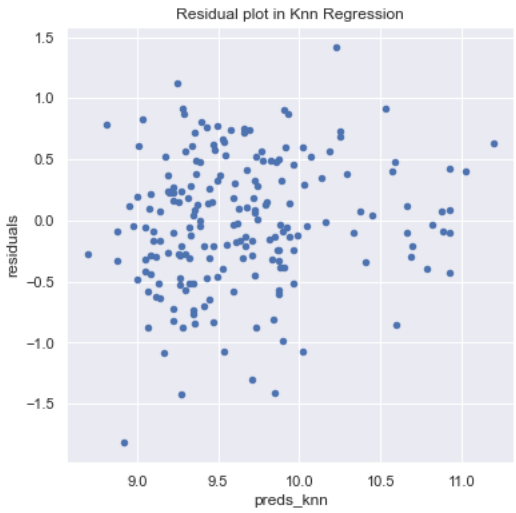


Fig 57 –Residual Jointplot (Knn Regression)

Fig 58 –Residual Jointplot (Lasso Regression)

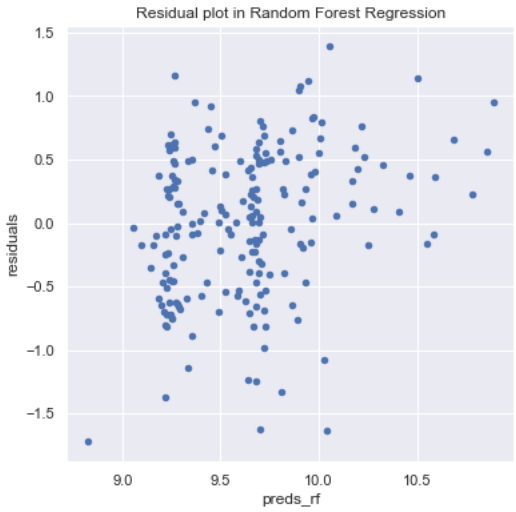


Fig 59 –Residual Jointplot (Random Forest Regression)

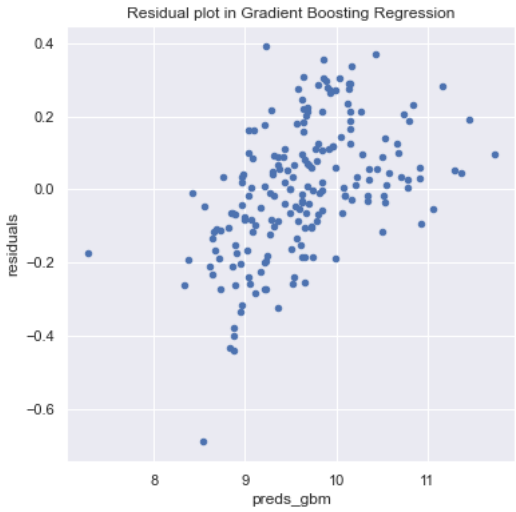


Fig 60 –Residual Jointplot (Gradient Boosting Regression)

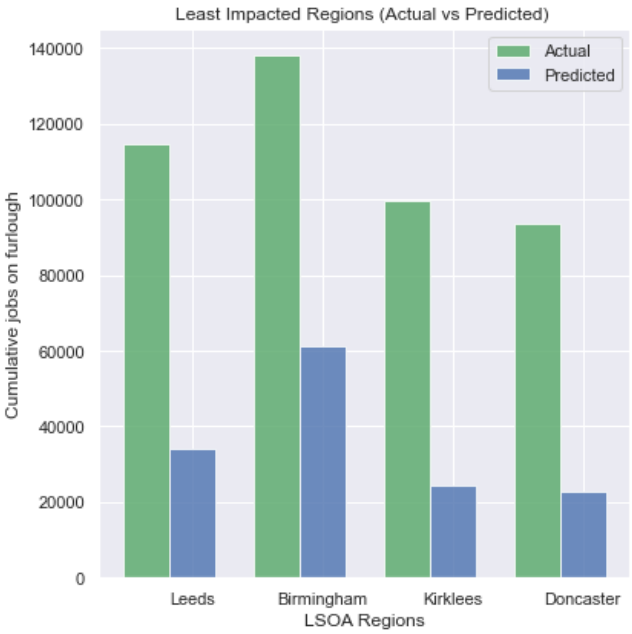


Fig 61– Least Impacted Regions on Furlough (Actual Vs Prediction)

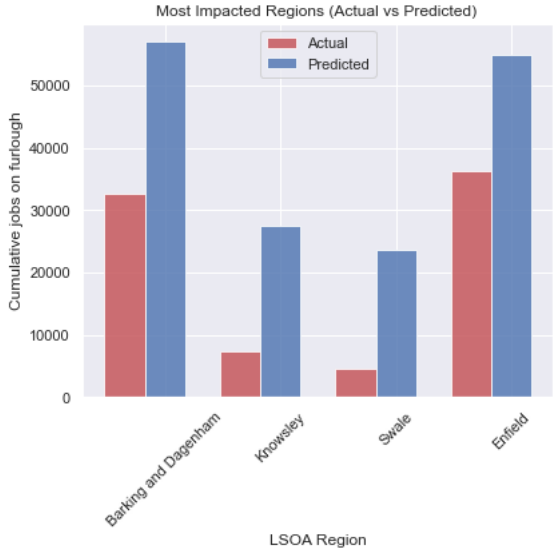


Fig 62 – Most Impacted Regions on Furlough (Actual Vs Prediction)

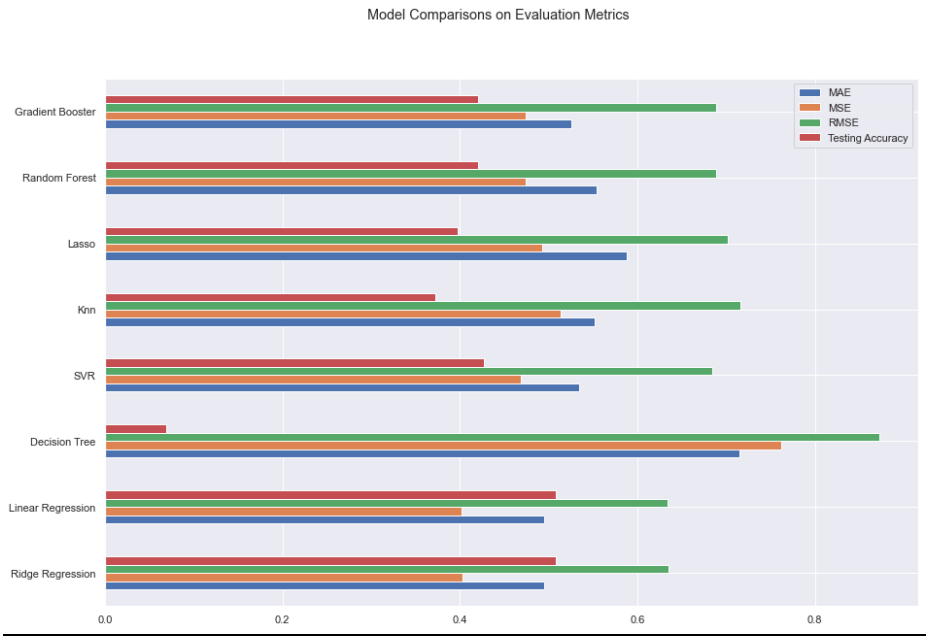


Fig 63 – Models Performance