Project Number 1
Sameer Nepal
DSC 680

Report on Project 1

Topic: This project predicts the number of people that will die because of heart disease. I will focus on the death count in one specific county within California.

Business Problem: Many people die every year because of heart disease, and this can be prevented if people take good care of their health. If government knows how many people are die each year, then they can invest time and money to early diagnosis of the disease. Predicting the number will give lawmakers overview of how bad the disease is and help motivating people by making some laws for mandatory checkups. According to the clinical practice guidelines for heart failure, it is recommended to utilize validated risk models to predict prognosis. This is essential in determining appropriate advanced treatments and hospice care by evaluating the accuracy of the models in identifying individuals who are likely to die within the next year [Appendix A].

Background/History: One in three adults in California are living with one of the heart diseases. The common forms of cardiovascular disease are heart disease, heart failure, stroke, and high blood pressure. There were 78,000 deaths in 2014 in California and LA county accounted for 15419 deaths. California Department of Public Health is actively engaged in various initiatives and endeavors aimed at enhancing cardiovascular health throughout the state. Their focus is on endorsing evidence-based programs that encourage healthy behavior and foster healthy communities. Predicting the number of deaths that might occur in the coming years will help in alerting the public about the health issue and creating healthy community.

Data Explanation: The data that I am using for my analysis is from California government data base and it has the information about the death count for different county and reason for death. The data were for range of different period, and I have imported two datasets for the date range of 1999-2013 and 2014- 2021. The data sets provide information on the yearly count of fatalities recorded in every region of California. The data set has been sub categorized by different methods like sex, religion and many more. I have performed the Exploratory data analysis to extract the data for LA county death every year caused by heart disease.

> Data link: https://data.ca.gov/dataset/death-profiles-by-county/resource/ca952a85-1c5f-43e0-aa8f-153eae30a0a5?inner_span=True
>
> Data Dictionary link: https://data.ca.gov/dataset/death-profiles-by-county/resource/54c8ec38-0ce7-4afc-8c39-7c19a9fefecb?inner_span=True

Two data set were merged to get the records from 1999-2021. The resulting datasets had 12 columns and around 400k records. The dependent variable for our analysis in the count of deaths and the independent variable is the year. Since I was only focusing on predicting the deaths in LA county, I did filter in the dataset to extract the data for LA county for the number of deaths that was caused by heart failure.

The scatter plot between year and number of deaths was studied and the pattern had some linear relationship. For the interval 1999-2013 the slope of the relation looked negative and for the interval of 2014-2021 the slope of the relation looked positive. Since there was linear relationship, I proceeded forward to find the model that would predict the number of deaths that might happen in future.

Methods: The main techniques I will be using is the linear regression analysis where count of death will be dependent variable and year will be independent variable.

Analysis: For the analysis of the data, I started with the scatter plot to see the relationship between the dependent and independent variable. I looked for any outliers using the box plot. Then I looked into the correlation between them. The correlation is not very strong, but I think this is okay because having perfect correlation is biased and training the model with highly correlated data might not be the good model. I divided the data into training and test sets in the ratio of 0.3 and fitted the data to the linear regression model. The model was generated and evaluated for its accuracy. The MSE and R squared for the model came out to be 643 and 45% respectively. R squared indicates how close the data is fitted to the regression line. Lower MSE and higher R^2 value gives us the better model.

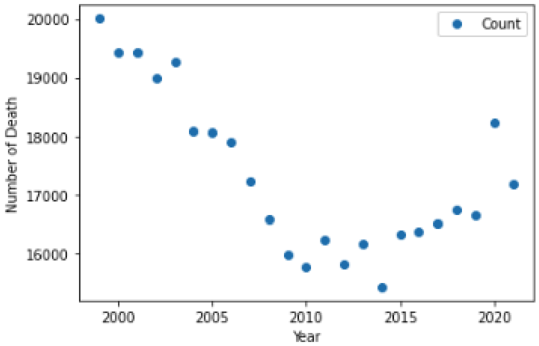

Fig 1: Scatter Plot

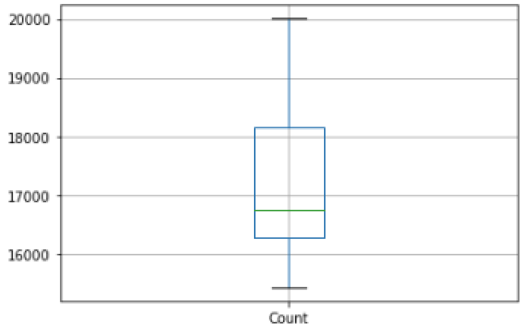

Fig 2: Box Plot

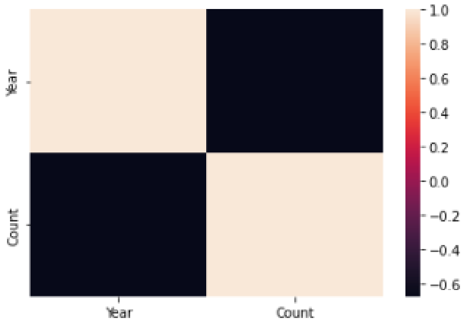

Fig 3: Heat Map

Conclusion: Using the model that we developed in this project we can predict the number of deaths that will happen in future. This will help health authorities in LA county to develop strategies on how this can be decreased by making the public aware of healthy habits. This might be the reason to further understand why the heart disease death number is increasing or decreasing in LA county. In addition, LA County can also share their experiences and lessons learned with other counties in the USA facing similar challenges. By collaborating with other counties, LA County can contribute to the development of evidence-based strategies to prevent deaths and improve health outcomes nationwide.

Assumptions: There are many assumptions that we have made while developing this model which is applicable to linear regression and some of the main one is as follows:

1. The linear relationship between the dependent and independent variables
2. The independent variables are not correlated with each other.
3. Linear regression model assumes that the error term is normally distributed.

Limitations: The model is very sensitive to the outliers. It is prone to under fitting where the accuracy of the model is very low.

Challenges: The data from 1999-2013 has the decreasing trend in number of deaths whereas the data from 2014-2021 has increasing trend in the number of deaths. This has skewed the distribution of the data when creating the model and hence the accuracy of the model is not very high.

Future Use and Recommendation: The accuracy of the model is about 45%. Since the model is generated using the actual data, I recommend using it in predicting the number of deaths in future. This will help the authorities in making necessary policies that will make public aware about the health of heart and its importance for life.

Implementation Plan: I will reach out to LA County officials and see if they will be interested in knowing more about this model and how it can help them make new policies that will help improve people's health from the cardiovascular side.

Ethical Considerations: The ethical concern that I need to be aware about is the private information of the people. Since the data is cumulative count of the people that have died it does not have personal information. It does not have any PII information other than the total number of deaths which are based on different categories like age group, disease, gender. This data is released by the state government, so the ethic has been considered while publishing the data.

References:

1. Gawali, S. (2022, July 22). *Linear regression in machine learning*. Analytics Vidhya. Retrieved April 2, 2023, from https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/

2. *Death profiles by County*. California Open Data. (2023, April 2). Retrieved April 2, 2023, from https://data.ca.gov/dataset/death-profiles-by-county

3. Centers for Disease Control and Prevention. (2022, February 25). *Stats of the states - heart disease mortality*. Centers for Disease Control and Prevention. Retrieved April 2, 2023, from https://www.cdc.gov/nchs/pressroom/sosmap/heart_disease_mortality/heart_disease.htm

4. Centers for Disease Control and Prevention. (2018, April 13). *Stats of the State of California*. Centers for Disease Control and Prevention. Retrieved April 2, 2023, from https://www.cdc.gov/nchs/pressroom/states/california/california.htm

5. Quantitative Finance & Algo Trading Blog by QuantInsti. (2023, March 16). *Linear regression: Assumptions and limitations*. Quantitative Finance & Algo Trading Blog by QuantInsti. Retrieved April 2, 2023, from https://blog.quantinsti.com/linear-regression-assumptions-limitations/

6. Larry A. Allen, M. D. (2017, April 1). *Predicting death among ambulatory patients with heart failure*. JAMA Cardiology. Retrieved April 2, 2023, from https://jamanetwork.com/journals/jamacardiology/fullarticle/2593745

Appendix A

Use of Risk Models to Predict Death in the Next Year Among Individual Ambulatory Patients With Heart Failure

The clinical practice guidelines for heart failure recommend the use of validated risk models to estimate prognosis. Understanding how well models identify individuals who will die in the next year informs decision making for advanced treatments and hospice.