Portfolio - Data Science Projects

Anuj Tanwar

Portfolio - Data Science Projects

Portfolio contains 6 Data Science projects that I completed during Masters of Data Science Degree.

**Note: Per requirement 10 projects were needed, however, I repeated same topics in multiple courses and kept on improving. So, I have submitted final version in the portfolio.**

**Portfolio Github Location:** https://github.com/anujtanwar12/Portfolio-Data-Science

Below are the projects:

1. **Temperature Prediction**

Weather is one that is not just close to us but is essential for our survival. Lot of businesses rely on weather, farmers rely on weather, bad weather can devastate the food on the fields. Sudden change in surface temperature can be harmful for our health as well. Research shows that abnormal weather disrupts the operating and financial performance of 70% of businesses worldwide. Every year, weather variability is estimated to cost $630 billion for the U.S. alone, or 3.5% of GDP. It becomes important to forecast weather in an accurate and timely fashion so that we can take the necessary precautions to minimize weather-associated risks. In the project I am predicting surface temperature using Long-Short Term Network (LSTM)-based model on more than 100 years of surface temperature recorded data from Kaggle.

Model prompts user to enter the city and search historical data of the city for current date of the month of each historical year. Then, it will clean the dataset and plot various graphs to understand the dataset. I will calculate mean, median, mode, Standard Deviation and Population Variance. I will use linear regression to predict future temperature.

**Building and Evaluating Model**

• Before building the predictive model, I prepped train and test datasets by using min max scaler that shrinks the data within the given range, usually of 0 to 1 and reshaping the datasets.

• Then I build a LSTM sequential predictive model with 4 neurons and 100 epochs. The mean squared error is being used as the loss function. Additionally, the adam optimizer is used, with training done over 100 epochs.

• Predictions were made on test data using model. Predict function to check the accuracy of the data.

GitHub Link: [Temperature Prediction](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/1.%20Temperature%20Prediction)

1. **Covid 19 Predictions**

India has seen an increase in COVID-19 case again during the start of year 2022. A simple google search can tell us that on April 4th, there were only 795 new cases, on Apr 14th there were 949 and in last week count of new cases have been on constant rise and going above 2000 per day now. This trend shows the cases have been rising. As per Ministry of Health and Family Welfare Government of India (https://www.mohfw.gov.in/) there had already been more than half a million documented deaths due to COVID and active cases are on rise. Government of India has also issued new guidelines and restrictions. Referring to the below google graph, we can see that COVID spread had increased in India during the summer months in last 2 years. 2022 summer season has already started in India, so the concerning questions that we have are:

* Are we going to have another wave of COVID?
* If so, then how severe it can be?
* When will we see the peak?
* How long did the previous waves last?
* What was the trend in death toll every day?
* What was the cured trend?

**Model building and evaluation**

* Model Used: Long short-term memory (LSTM). It is an artificial neural network used in the fields of artificial intelligence and deep learning.
* I have used sequential model from tensorflow.keras package.
* Added a long short-term memory layer with 100 memory units
* Used rectified linear activation function (RELU) which will output the input directly if it is positive, otherwise, it will output zero.
* Used 20% dropout.
* Compiled the model with adam optimizer

**Github Link:** [Covid-19 Predictions](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/2.%20COVID-19%20Predictions)

1. **Movie Recommendation System**

Recommendation Systems play a crucial and important role in this “era of abundance”. For any product, there are thousands of options. Recommendation systems help to personalize a platform and help the user find something they might like. Online streaming services such as Netflix, Disney, YouTube, Amazon, etc use these recommendation systems to make the right recommendations about the movies and shows can have drastic impact on the customer base and the overall business. If customers are satisfied with the recommendations, they will spend more time using the service and would not look for other streaming services options which leads to customer retention and growth in sales and profit. Various sources say that as much as 35–40% of tech giants’ revenue comes from recommendations alone. The Netflix offered a Prize of 1 million dollar and hosted an open competition for the best recommendation algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

**Modeling and Analysis**

I am using Netfilx data and MovieLense data to demonstrate my approach to address the problem. Below are the steps involved with Netflix data:

* Download the datasets
* Read the dataset files in different dataframes
* Clean the data
* Transform the data to have customer and movies related data along with Ratings given to the movies.
* Now we can use this dataframe to do the regression and predictions

**Github Link:** [Movie Recommendation System](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/3.%20Movie%20Recommendation%20System)

1. **Changing Fashion Trends with Time**

Fashion has always been a changing and evolving field. Fashion industry is going through huge innovations with changing Technology. Fashion is observed in anything a person has or does. Ranging from clothes, shoes, watches, TVs to cars, food and houses.

Project involve researching online on how fashion trends are changing and factors influencing it.

**Github Link:** [Changing Fashion Trends](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/4.%20Changing%20Fashion%20Trends%20with%20Time)

1. **Weather Prediction**

This is a Project for Python beginners that includes forecasting weather using Open Weather Map API.

**Github Link:** [Weather Prediction - API](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/5.%20Weather%20Prediction)

1. **Airlines Data Analysis**

Since the disappearance of MH370 in 2014 and two Boeing 737 crashes, first in 2018 and second just 5 months later, there has been fear in public and concerns about the air travel safety. However, despite this increased attention we’ve seen in media reports, these numbers do not really represent any increasing trends. On the contrary, we’ve seen the opposite trend as airline accidents have continued to decrease over time.

I have chosen Power BI for creation of below data visualizations:

1. **Donut Chart**

I have divided the period between two groups, 1985-1999 and 2000-2014, and then compared the Air Traffic Fatalities between them. This shows, how fatalities have reduced almost to half during 2000-2014 as compared to 1985-1999. Hence, showing how the air traffic has become more safer over the period of time.

1. **Line Chart - Fatalities by Year**

This chart shows total Number of Fatalities from 1920 to 2020. From the plot it is evident how Air Fatalities have reduced from 5k in 1940 to 0.5k in 2020. I have created a filter to filter out years between 1920 and 2020 as there is insignificantly less data for rest of the years.

1. **Bar Chart**

Air Traffic by Year This chart shows total air traffic from 1970 to 2020. From the plot it is evident how Air Fatalities have reduced significantly in last 50 years. From above line chart and this bar chart, it is evident that Fatalities have reduced when air traffic has increased multiple folds over the period of time which in term depicts that air traffic has become safer.

1. **Line Chart – Passenger & Revenue**

This chart is showing the Passenger count as well as Revenue as two different lines over the same period range. This shows how the passengers and revenues have shown to grow over the period and air travel has become more and more popular.

1. **Horizontal Bar chart – Fatalities by Each Airline**

Comparison of Total Fatalities by group of years, 1985-1999 and 2000-2014, for each Airline. Chose bar chart because it shows best comparison between two groups for each airline. For example, Malaysian Airlines had more fatalities between 2000-2014 as compared to 1985-1999.

1. **Horizontal bar chart - Available Seat KM Flown Every Week (Billions) by Airline**

This bar chat shows the sheer volume airlines experienced on a weekly basis.

**Note**: This project is still under development and will complete by the end of course.

**Github Link:** [Airlines Data Analysis](https://github.com/anujtanwar12/Portfolio-Data-Science/tree/main/6.%20Airlines%20Data%20Analysis)