Ireland Tourism Prediction

Group 4

Amogh Anil Rao (22306378) Ketan Patil (22303876) Karan Dua (21331390)

Introduction:

Tourism is a booming industry in Ireland and is a source of income generation for the locals. Predicting the arrival of tourists is a complex task as it involves multiple factors and many of these factors are volatile in nature. Therefore, we find this subject an interesting area to research and build solutions, where we would provide expected tourist arrivals from different countries. We are trying to analyse if the parameters like search interest of people visiting Ireland (Google Search Trends from various countries), Weather Conditions in Ireland, and people visiting from specific countries (Historic country-wise tourist arrival data) have an impact on the overall tourist arrival numbers in Ireland.

To make our machine learning model, we have tried to incorporate different features in the data that may have an overall impact on the tourist arrival in Ireland and tried to analyse the impact of these features on the data.

Dataset and Features:

To make our machine learning model that can predict the arrival of tourists in Ireland, we must gather the following data:

1. Weather data:

To gather weather data for Dublin, the capital of Ireland, we have done Web-scrapping from Wikipedia(https://en.wikipedia.org/wiki/Dublin) using Beautiful-Soup. Once the data is scrapped, we performed formatting on the data to remove special characters and formatted the data to valid datatypes. The temperature data contained Celsius and Fahrenheit values in one cell, we have removed the Fahrenheit value and taken the Celsius value into consideration. Each value stored in a string is then converted to a floating-point value so that it can be processed accordingly later. The data was then mapped to months of the year. Once we formatted the data, we had the following columns in the dataframe:

- Month: Signifies the month when temperature readings were taken
- average_high_celsius: Signifies average high temperature for all days in Dublin during a month
- average_low_celsius: Signifies average low temperature for all days in Dublin during a month
- average_precipitation_days: Signifies on average how many days it rained in Dublin during a month
- average_precipitation_mm: Signifies on average how much it rained in Dublin during a month
- average_relative_hummidity: Signifies average humidity for all days in Dublin during a month
- daily_mean_celsius: Signifies average temperature for all days in Dublin during a month

2. Historic tourist arrival data from different countries in Ireland:

To get historic tourist arrival data, we used the CSO website(https://data.cso.ie/table/ASM02). This dataset contained the following features:

- Statistic – This column contains a static value (Air and Sea Travel)
- Month – It contains the year and month in which tourists arrived in Ireland
- Country – It signifies the country from which tourists arrived in Dublin in that month

- Direction – This column contains a static value (Arrival)
- UNIT – This column contains a static value (Thousand)
- VALUE – It contains the number of tourists that arrived from a country in a particular month.

Based on the above data, we have first removed the column Statistic, Direction, and UNIT from the data as these contained static values. Next, we multiplied the VALUE column by 1000 to get the actual number of tourists arriving in Dublin. Furthermore, the Month column contained year and month merged together. To get unique features from this, we sliced this column to get two separate features Year and Month.

Now, this dataset also contained some rows where unique country names were not provided like Other Countries(42), Other Europe (34), Selected EU (AT, BG, CY, CZ, DK, EE, FI, GR, HR, HU, LT, LU, LV, MT, RO, SE, SI, SK), Other Transatlantic Countries(1), We have removed these rows from the data as it was not possible to get google trends data for merged countries.

Also, we renamed a few countries to map them with the Google Trends data.

3. Google Search Trends data

Based on the countries and years that were extracted from the above data, we have manually downloaded Google trends data. For instance to find the search interest for Ireland Tourism by people in Belgium use the following URL: https://trends.google.com/trends/explore?cat=208&date=2010-01-01%202022-12-02&geo=BE&q=ireland.

Once the data is downloaded for all 12 countries for the year 2010 to 2022, we processed this data to replace NaN and 0 values with backward filling and forward filling approaches. In this approach, the missing value is replaced with the value appearing row before or row after. Also, we changed the datatypes of the column for further processing. Once we formatted the data, we had the following columns in the dataframe:

- Month
- Year
- Country
- VALUE

Now, we have 3 different datasets, and we must merge them. We first merged Dublin temperature data with tourist arrival data using the inner join on the Month column of both datasets. Then, we merged this dataset with Google Trends data again using the inner join on the Month column.

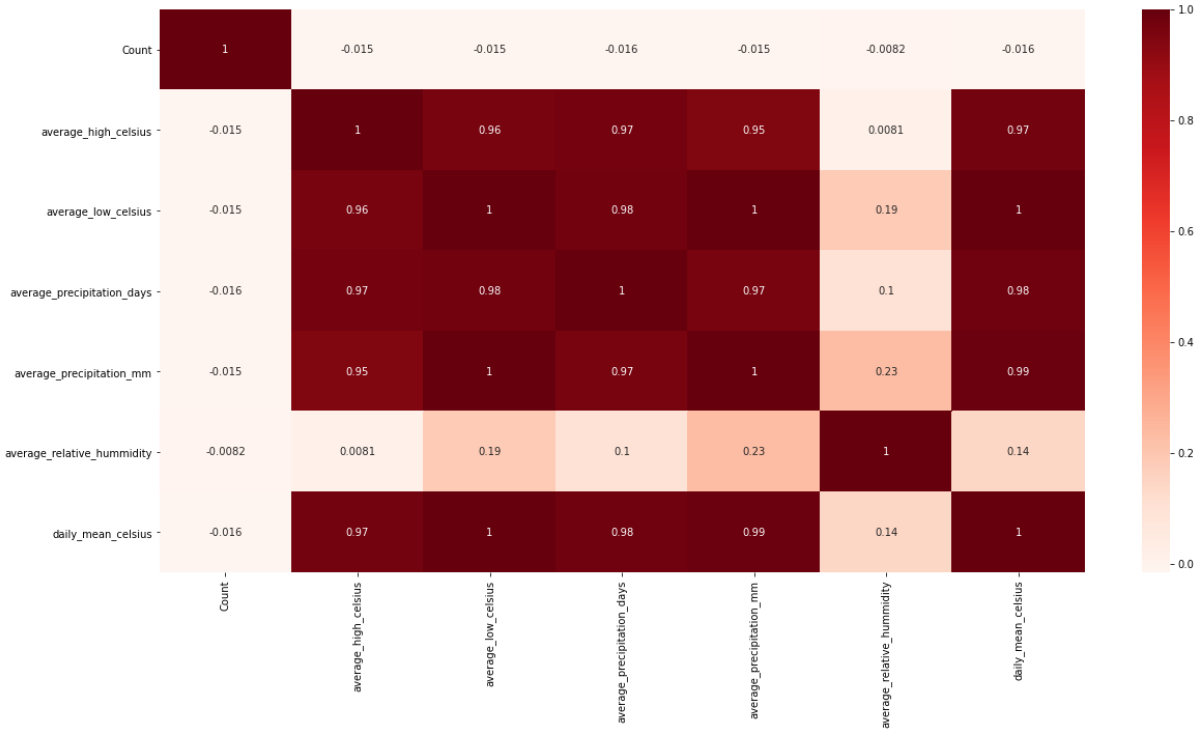
This gives us a complete dataset for our analysis.

For the Feature Selection process, we first removed the Value column from the dataframe as it was the value to be predicted by the model. We saved this column into a CSV so it can be used later during model training. Then, we removed the categorical variables from the dataset. For the remaining Quantitative data, we generated the heat map to get the correlation between various features. The following heatmap was generated:

Ireland Tourism Prediction

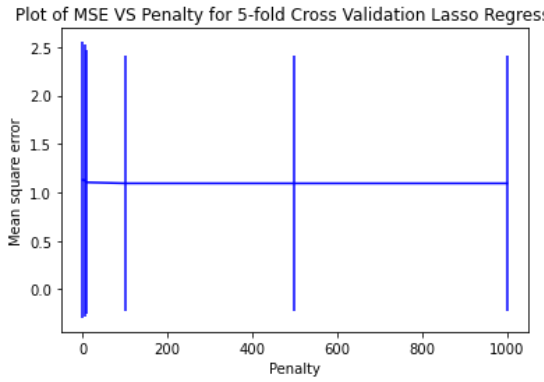
Group 4

Amogh Anil Rao (22306378) Ketan Patil (22303876) Karan Dua (21331390)

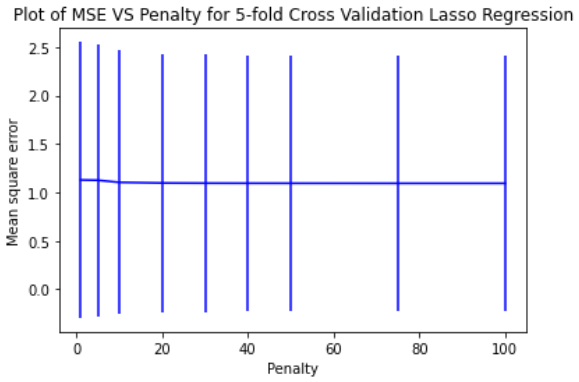


Based on this heat map, we can analyse that average_high_celsius, average_low_celsius, average_precipitation_days, average_precipitation_mm, average_relative_hummidity, daily_mean_celsius are highly correlated.

Highly correlated variables do not add additional information to the model but rather increase the complexity of the model and increase the chances of errors. Therefore, we decided to remove these features. To choose which features to remove we used Lasso Regression with L1 regularisation. We decided to use Lasso Regression as this model reduces the insignificant coefficients in the model to zero. To train the Lasso Model, we first normalised the data using StandardScaler library. Then, we created a method that performs K-Fold Cross validation on the data and returns the mean of Mean Square Error. We tested different ranges of hyperparameters and generated plots for Mean Square Error for different Penalty parameters.



For C = [1, 5, 10, 100, 500, 1000]



For C= [1, 5, 10, 20, 30, 40, 50, 75, 100]

Ireland Tourism Prediction

Group 4

Amogh Anil Rao (22306378) Ketan Patil (22303876) Karan Dua (21331390)

Firstly, we created plot on broad range of penalty parameters starting from 1 to 1000. By analysing this plot, we analysed Mean Square Error has converged for very small value of C. Then, we created plot for penalty parameter 1 to 100. By analysing this plot, we identified the elbow like structure. Using 'Elbow Method' we identified that C = 10 is ideal point in graph where MSE does not decrease significantly for higher values of C.

Using C = 10, we created Lasso Model and got below result and we were able to identify the insignificant features in the dataset by finding the coefficients that were reduced to zero.

	Penalty	Intercept	Count	average_high_celsius	average_low_celsius	average_precipitation_days	average_precipitation_mm	average_relative_hummidity	daily_mean_celsius
0	10	[[0.079]]	[0.222]	[0.012]	[0.079]	[0.0]	[0.0]	[0.0]	[0.019]

Based on above result, we removed the following features from the dataset: average_precipitation_days, average_relative_hummidity and average_precipitation_mm.

We have also removed the year column from the data to remove as we are only predicting month-wise tourist arrival in Ireland so it makes this parameter insignificant for prediction and it can lead to overfitting in the model.

For Feature Engineering, we used one-hot encoding for the categorical variables in the dataset like Month and Country. One-hot encoding encodes categorical data that to ordinal data. For instance, each country is created as a new feature and the value of 1 is given when it is applicable. This is done to convert the String data to Numeric data for increasing the machine's understanding of the data as Machine Learning models cannot understand the alphanumeric values. E.g., In Country Column, Belgium is one of the values, for this Country_Belgium feature gets added and a value of 1 is given when the corresponding row value is for Belgium Country and the values for the rest countries are 0.

We saved these final features set after Feature Selection and Feature Engineering Process into a CSV.

We have tried to find highly correlated features in the temperature data. On the found correlated data we have plotted a seaborn heat map to visualize the correlation between various features. Based on the plotted graph, out of the highly correlated data we have chosen only one of them as it will not add additional information to the model but rather increase the complexity of the model and increase the chances of errors.

Model Selection and Results Discussion:

Baseline Model: Linear Regression Model

Linear regression is used in predicting the value of a variable based on the value of another variable. It finds a line that best fits the data points. It is used to evaluate trends and give estimates or forecasts.

Linear regression checks if the input variables are good enough to estimate the output and which input variables in particular act as significant predictors. They show the relationship between the dependent variable (expected outcome) and the independent variable (given input). We chose the Linear Regression model as the Baseline model because it is a very simple model, easy to interpret and no hyperparameter tuning is required in this model.

We have also used a 5-fold cross-validation score as a benchmark score from this model for the evaluation of further models. The primary reason for this is that the cross-validation score reduces the effects of

overfitting and underfitting in the model as every part of the data is used to train the model so the model has more chance to generalise on the data. This helps the model to adapt better to unseen test data. Also, the cross-validation averages out the error of the trained model to reduce the impact of the shuffling of data. Furthermore, we have used Mean Square error as a scoring strategy for cross-validation score as it is an ideal score for Regression as MSE explains how close the predicted value was to the actual value. Therefore, a smaller MSE value is considered ideal for any regression model.

For Baseline Linear Regression Model, the Cross Validation Score (absolute of negative MSE) is: 0.2

This is the benchmark score for our evaluation of further models.

Model 1: Kernelized Support Vector Regressor

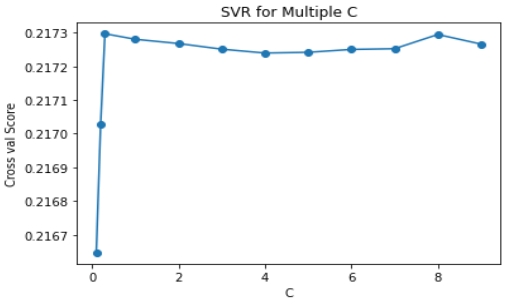
Support Vector Regressor (SVR) is a supervised learning algorithm. The key idea is to find a plane that captures the maximum points in the data. The more points that fit this plane, the lower the error in predicting the values as the model can find a generalised plane for the data. This is generally used to predict discrete values. SVR allows us to define the total allowance of the error to be allowed in the model.

We choose this model for our regression problem as SVM tries to adapt to variations in the features using different kernels. This can help to generalise model parameters and prevent the model from overfitting.

There are 2 hyperparameters that should be tuned for the SVR model, namely kernel and regularisation parameter C. Kernel is used to define the shape of the plane. The linear kernel generates a linear plane to capture the data while RBF is a radial kernel and is generally used if there is non-linearity in the data.
We again used Cross validation Score with an MSE scoring strategy and tried different variations of the SVR model to tune the hyperparameters.

1. Results for Linear Kernel SVR with different Regularisation Parameter:

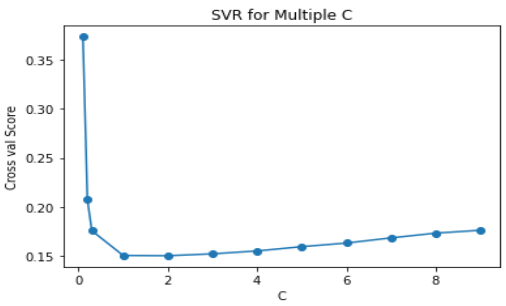
```
Cross Validation Score for SVR is: 0.21664676634816252 C =  0.1
Cross Validation Score for SVR is: 0.21702644394257448 C =  0.2
Cross Validation Score for SVR is: 0.21729675613258972 C =  0.3
Cross Validation Score for SVR is: 0.21728031373328965 C =  1
Cross Validation Score for SVR is: 0.21726770848925053 C =  2
Cross Validation Score for SVR is: 0.21725095139745138 C =  3
Cross Validation Score for SVR is: 0.2172394511973193 C =  4
Cross Validation Score for SVR is: 0.21724160336428802 C =  5
Cross Validation Score for SVR is: 0.2172501373326165 C =  6
Cross Validation Score for SVR is: 0.21725228205623331 C =  7
Cross Validation Score for SVR is: 0.21729445784982082 C =  8
Cross Validation Score for SVR is: 0.21726598409561865 C =  9
```



Based on the above results the best regularisation parameter for Linear SVR is 0.1 as absolute MSE is minimum for this value of C.

2. Results for Linear Kernel SVR with different Regularisation Parameter:

```
Cross Validation Score for SVR is: 0.3732999538221743 C =  0.1
Cross Validation Score for SVR is: 0.20727533567990367 C =  0.2
Cross Validation Score for SVR is: 0.17571664232080675 C =  0.3
Cross Validation Score for SVR is: 0.15049333711263252 C =  1
Cross Validation Score for SVR is: 0.1502210516871648 C =  2
Cross Validation Score for SVR is: 0.15214800132433212 C =  3
Cross Validation Score for SVR is: 0.15509909940524053 C =  4
Cross Validation Score for SVR is: 0.15944903720699108 C =  5
Cross Validation Score for SVR is: 0.16313639221953083 C =  6
Cross Validation Score for SVR is: 0.16840763208429438 C =  7
Cross Validation Score for SVR is: 0.1732216294850756 C =  8
Cross Validation Score for SVR is: 0.17617847455349744 C =  9
```



Based on the above results the best regularisation parameter for Radial SVR is 2 as absolute MSE is minimum for this value of C.
Based on the above results, we can conclude that the SVR model with Radial Kernel and Regularization parameter values as 2 is best for this data. The primary reason for this is that our features are non-linear and are well captured by the radial kernel.

Tuned SVR model Cross Validation Score (absolute of negative MSE): 0.1502

Model 2: Decision Tree Regressor

The decision tree algorithm is a supervised machine learning algorithm that divides the dataset into smaller subsets which further divides the subsets till a leaf node (data set which cannot be divided - Molecular dataset) is generated. Ideally, they have 2 or more nodes with the root node (topmost node) being the best predictor and the internal nodes being the features of the data set.

We choose this model for our problem statement as decision trees can accurately adapt to regression problems and identify and split the tree nodes on the basis of the importance of the feature in the overall prediction of the tourist arrival.

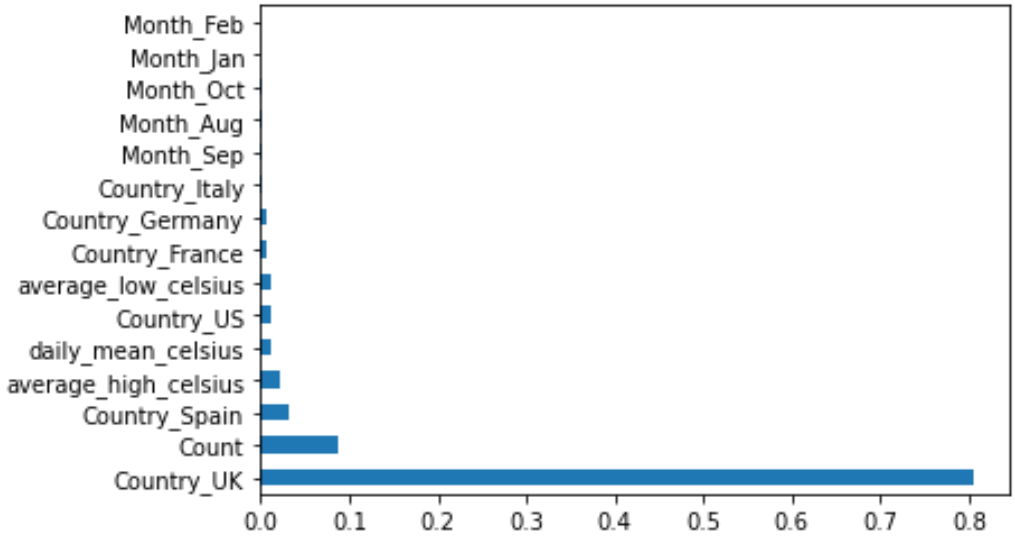
For this model, there are 3 hyperparameter that needs to be tuned. The Splitter defines whether to use the best feature or random feature to split the tree, Max_Depth defined the maximum depth of the decision tree and we have provided a range of values for which this model has to be tuned, and Max_Features defines the number of features to be included in the split.

Based on our hyper parameter tuning for decision tree regressor, we have got the below result:

```
The best Params for the decision tree  {'max_depth': 6, 'max_features': 'auto', 'splitter': 'best'}
```

Now, we have trained the decision tree regressor based on the above results and we obtained the **Cross-validation score using the absolute of negative MSE strategy as: 0.1627**

Also, we identified the feature importance (Top 15 Features) for this model:



Based on the above graph, we can conclude that most tourists that travel to Ireland are from the UK hence it is the most important feature, and we can also analyse that Google Search Interest is the second most important factor that impacts Tourist arrival. Furthermore, temperature parameters like average_high_celsius and daily_mean_celsius are also important in the model.

Ireland Tourism Prediction

Group 4

Amogh Anil Rao (22306378) Ketan Patil (22303876) Karan Dua (21331390)

Summary:

Based on the above experimentation we can conclude that both the models that we trained have better performance than the benchmark baseline model. Also, the SVR model performs best on this data as it has the lowest MSE. The main reason for this result is that the SVR model was able to capture the non-linearity of the data and adapt well to this data. This led to a better generalisation of the model which other models could not.

Also, we were able to identify the Top 15 features that plays important role in overall tourist arrival prediction in Ireland.

Contributions:

1. Data Gathering: Karan, Amogh
2. Data Pre-processing: Karan, Ketan
3. Feature Selection and Feature Engineering: Ketan, Amogh
4. Baseline Model: Karan, Ketan
5. SVR Model: Ketan, Amogh
6. Decision Tree Model: Karan, Amogh
7. Report Writing: Karan, Amogh and Ketan

GitHub Repository :

Repo URL: https://github.com/Amogh4u/ML_Ireland_tourism_prediction