National College of Ireland

Auto-Recommendation of Movies from Weather Prediction using Machine Learning: Dublin, Ireland

MSc Research Project
Data Analytics

Kanak Kaushik
Student ID: 18136966

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kanak Kaushik
Student ID:	X18136966
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Christina Muntean
Submission Due Date:	12/12/2019
Project Title:	Auto-Recommendation of Movies from Weather Prediction using Machine Learning: Dublin, Ireland
Word Count:	6360
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not enough to keep a copy on the computer.	

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Auto-Recommendation of Movies from Weather Prediction using Machine Learning: Dublin, Ireland

Kanak Kaushik
X18136966

Abstract

With new movies releasing each week, the film database is developing hastily and the deciding on a movie from the lot is turning into greater difficult than ever. This has caused researches achieved in this field to instigating services of recommending a film to make a higher experience for user experience. Therefore, this research recommends a concept to mass users with respect to film and video streaming offerings. A version is proposed with the intention to automatically recommend a movie influenced by the aid of the weather. ARIMA forecasting model is used for weather prediction with RMSE of 0.298 and Random Forest was used for classification of the movies with 57% accuracy affected by the weather.

Keywords: *Music and Video streaming services, Weather prediction, KNN, UX, ARIMA, ES, Random Forest*

1. Introduction

First-ever a film made turned into in 1878, through Eadweard Muybridge to reply to a scientific question: Are all the legs of horses are in the air even as galloping? Thus, moving pictures (movies) become born. This medium allowed us to transfer and understand data visually. It developed to illustrate the artwork of storytelling to larger audiences at the same time and maybe dispensed globally. People adopted this concept without difficulty and quick and began experimenting with it. Nowadays, 'Movies' are a major source of enjoyment and its industry has been growing gradually. It allows human beings to enjoy visible sensationalism, lovely imagery, gripping soundtracks, and colorful dialogues (Nakatsu and Edirisinghe, 2011). Diverse subculture humans have diverse flavors in the films. The movie industry now has various productions homes, groups, trades, and lots of extras.

Ireland has emerged as a hotspot for filmmaking, due to its attractive manufacturing environments like moneymaking tax incentives, phase 481 (tax credit incentive), and a plethora of extraordinary writers, directors, actors, and team. A couple of films like 'Brooklyn' and 'Room' become nominated for Academy awards lower back in 2016 (About the Irish Film Industry beauty and satisfied-pass-fortunate humans filming movies in Ireland is becoming not unusual. Movies are getting more influence on the audience and thus, brands are seeing this as an opportunity to tap in, by getting signing deals to promote their products in the movies, which will help them to boost their consumer base(Chavadi, Menon and Sirothiya, 2019).

Loads of movies are releasing on the box office every weekend, also digitally releasing on the video streaming services. Because of this, users are having plenty of options to choose from the lot. But this creates a snag, of what movie to watch. Streaming services like YouTube,

Netflix, Amazon Prime Video, and many more are playing a pivotal role in supplying the end-users demand by giving them options to choose from. Every user reacts differently in various environment, this is called context-awareness situations (Kaminskas and Ricci, 2012). This area of recommending content to the end-user depending on their reactions is on the rise. Even movies can be affected by various regions it is premiering in, cultural values of certain communities have also effect on which movies will be well-received (Sinha *et al.*, 2019).

Services like Netflix, Amazon Prime Video, Sky Go and IMDb offer film suggestions prompted by way of the customers watching behavior, clicks, customizable playlists, and several other. Netflix, a film and TV display streaming offerings, based in 1997, is one of the main players in this enterprise. It has accrued astounding 118 million subscribers in 2019. Although it initially commenced as supplying rental as well as on-demand movie offerings, now it has its own manufacturing house which is making nicely acclaimed films within the business (Cook, Sam, 2019). To get to this status, Netflix has carried out diverse techniques for the offered offerings to make the stop-person revel in excellent. Thus, this studies venture alludes to one such concept to complement the consumer experience.

Human mood usually is affected by the surrounding atmosphere, which then affects the choices made by humans for choosing a movie. The weather has been observed by many researchers to have an effect on human life (Senior Psychometric Analyst — Mettl. Bangalore *et al.*, 2019). Different movie genres represent different tastes, like whilst there are wet days human beings like to observe romantic and tender movies whereas, the sunny day represents a more active, pleased, and upbeat film genre.

The goal of this project is to try and instigate the relationship between movie and weather, to categorize movies based on climate conditions and as a consequence recommend movies. It will help to make the user experience easier.

1.1 Research Question

This study tries to find the answer to the following research question:
"*How accurately and efficiently Machine learning algorithms help to recommend movies to the user based on the weather conditions?"*

2. Related Work

2.1 Introduction:

The quantity of data has been growing as the days are going. The researchers thought of consuming this data inside the field of analytics to get a few insights about the marketplace, for you too, in flip, assist the agencies to develop. Data analytics is the technological know-how wherein the raw information is analyzed to make a few conclusions from the statistics. Many techniques like prediction, recommendation, forecasting, junk mail observation, and lots of more come beneath the sphere of statistics analytics. These techniques can display tendencies and metrics that could in any other case be difficult to get from the heaps of facts. There are diverse sorts of statistics analytics: Descriptive analytics, Diagnostic analytics, Predictive analytics and Prescriptive analytics (Frankenfield, 2019). This study will recognition on Predictive analytics, which means what is going to take place inside the future from the formerly determined facts. Companies take gain of information analytics in a suitable manner for his or her enterprise. Business intelligence structures, Decision support

systems, in addition to using one-of-a-kind techniques of the system getting to know and records mining technologies is involved in information analytics (Department of Information Technology et al., 2019).

Machine learning also referred to as predictive analysis, or predictive modeling is a sub-kind of Artificial Intelligence. It makes use of programmed algorithms for the analysis of the person enter input and then to expect the output, they analyze and optimize themselves to provide better outputs. There are 4 sorts of machine studying algorithms, this study might be the use of a supervised learning algorithm, to classify, regression, and forecast. People were predicting climate from the early 18th century. This has been limited to its programs and has now not been accurate. But for this, machine learning algorithms were evolved to forecast the climate more appropriately than ever and constantly improving itself. The climate forecast is being implemented in many fields like tourism, farming, transportation, and much greater.

A case-based reasoning system is a discipline beneath artificial intelligence which deals with shrewd reuse of know-how (results/consequences) from already solved to advantage a few insights for the brand new and but unsolved problem. The kernel of the research done by using Winoto and Tang is to retrieve and classify the genres of films primarily based on the user temper, or other factors main in changes to the choice of a particular film (Winoto and Tang, 2010).

2.2 Related work:

This part will shed some light on the field of prediction that had to affect the weather conditions of the place. Many types of research have done the study on and presented their findings.

2.2.1 Prediction using weather data:

Forecasting the weather and the use of it because the input variable for many different selection factors has been used previously. Certain researches were studied, and their pivotal findings had been pronounced in aid of this study.

2.2.1.1 Energy

De Felice, Alessandri, and Ruti, (2013) have used numerical weather prediction (NWP) models to forecast the climate a good way to predict energy loads in a region in Italy. The authors have taken relationships between power and temperature, some of the cooling pieces of equipment over a place, irrigation at some point of summertime and heating all through winters. These factors have an effect on the consumption of electricity over the seasons. Naïve predictor and ARIMA/X models had been used for forecasting the seasonal weather trend. The data for energy becomes sampled to common use within the day and become divided into 9 grids depending at the place. They applied LOESS (locally weighted polynomial regression) at the dataset to fit the data and then plotted the differences in graphs. And for the climate records mean of all of the statistics factors have been taken to run models on the dataset. The authors have as compared and analyzed the outcomes from all the fashions and have determined to go along with ARIMA/X and NWP fashions which may be used on this study's technique. A different study became done through (Moazami et al., 2019) for the city of Geneva, the effect of weather on buildings' energy performance. They downsized the records documents into 3 groups: typical downsized yr, extremely hot year and extremely cold year and carried out algorithms on it to get the effects. They confirmed that the climate does have an effect on the electricity performance of a building with low-to-no insulation in the walls.

2.2.1.2 Travel

Travel time to reach any destination depends on the congestion ahead on the road, but these factors aren't simplest the only that reasons it, but factors like weather, time of the day, and lots of more also make contributions to travel time. There turned into a take a look at carried out by means of Kamga and Yazıcı, (2014) about the factors affecting the tour time for New York City. They took GPS and weather datasets for their studies. They carried out statistical analysis in calculating journey time like common travel time, standard deviation, and coefficients of variance for the studies. They took only travel time from the dataset to calculate it, then did a descriptive evaluation at the dataset. They used type and regression tree models on the dataset for the research. The authors took weather into play whilst the present process this model, they determined that on snowy days and the wet day's tour time can be higher than regular. J. Javid and Jahanbakhsh Javid, (2018) additionally did a have a look at travel time in California. They took the same parameters for the records however alternatively ran a robust regression model on the dataset. The authors suggested that robust regression was better than the normal regression method because it weighs different elements too. The authors concluded that with the version proposed they could enhance the overall journey time through almost 60% through imposing the method. But this took not varied data for the model which can arguable, as this regression version could work at the given dataset and could work in another way at the others.

2.2.1.3 Accident

Road accidents have a negative effect on each individual and societal harm. Malin, Norros and Innamaa, (2019) did a have a look at on the risks of an accident on ways because of terrible climate situations on Finnish roads. The statistics become accrued from police-pronounced injuries and the weather data from Foreca Ltd. They used Palm opportunity, which means that that the occasion takes place indefinitely. They correspond it to choosing any random vehicle from the traffic. They executed statistical analysis at the datasets to get the outcomes. They found that climate is certainly an essential component within the number of accidents. As the roads have sleet or snow or moist roads it impacts the grip of the tires of the automobile to grip on the street. Palm chance is a good approach to apply. Another look was completed by using Cheng et al., (2017) within the consequences of time and weather on crash types. In this, the authors used the Bayesian and Multivariate Poisson lognormal (MVPLN) model to decide the effects. The findings were in comparison within the table and determined the MVPLN model to be top for the dataset.

2.2.1.4 Environment

Ocean surface currents were additionally predicted via the weather records. Kalinić et al., (2017) accomplished on this research. The training and testing datasets have been used from HF oceanographic radar. Numerical weather prediction (NWP) and neural network self-organizing map (SOM) became used to perform the studies. SOM was used to visualize complicated statistics and create abstractions like in other classification algorithms. It is an unsupervised learning algorithm. And NWP was used to predict the numerical weather records. But SOM has a downside that its excellent of results is as correct because the input that is being given to the model.

Similarly, Zhang et al., (2019) did an examine in milk manufacturing forecasts models affected by the climate. The authors wanted to analyze the impact of introducing a few climate information to the milk manufacturing forecast models with the help of a Non-linear Auto-regressive model with Exogeneous input (NARX) and Multiple Linear

Regression (MLR) models. A descriptive analysis was done on the dataset. From the research, the authors concluded that the NARX model becomes higher than MLR due to its better R2, RMSE and SSE values mentioned with the aid of the software. But the effects were not strong enough to say that by means of introducing climate aspect the prediction of milk production had been affected.

Thus, all the last researches concluded that weather does have an impact on different predictions of the fields and can be used in many more such use-cases.

2.2.2 Case-based Reasoning application:

Case-based reasoning is a machine learning algorithm that uses past results to predict future results. It is used in a variety of fields like recommendation systems, prediction systems, classification systems, and many more. CBR is a sub-type of a machine learning algorithm (Pla *et al.*, 2014). Boral, Chaturvedi, and Naikan, (2019) conducted a study on fault detection and isolation system with the help of case-based reasoning. The authors proposed a method for faulty gear detection, as it will help inexperienced engineers to easily detect faults and solve it. It will help the engineers to be more practical with their approach. Another study was done by Basu, Roy, and DasBit, (2019) for post-disaster resource planning. The authors have used case-based reasoning and Principle component regression analysis (PCRA) model to predict the resources which will be in demand after the disaster. The results from PCRA were validated from the CBR system. CBR uses K-nearest-neighbour (KNN) algorithm for the classification of the results. These studies show that CBR is excellent when there are problems that have a high probability of being influenced by old data.

2.2.3 ARIMA and ES

Auto-Regressive Integrated Moving Average (ARIMA) has been used for time series forecasting repeatedly. ARIMA model was used by Mitkov *et al.*, (2019) in forecasting the energy consumption over Afghanistan. The authors found the ARIMA (1,1,1) model to be best for their study. A graph showing original vs modeled approach was put there and from that, it was concluded ARIMA worked well. It is also implemented in the prediction of Fuel cost Zhao *et al.*, (2018). The proposed model was superior to the previous three-month delayed approach.

ES (Exponential Smoothing) is used for short time series prediction. Sahay et al., (2018) used ES for the detection of black hole attacks in the IoT environment. It was used to estimate the time of the attack and detect the malicious node and remove the node from the network. It was also used for Bus time travel and arrival time prediction. As ES is a bit easy from the ARIMA model, it was used with Kalman filtering to update the data, as time goes on. The result found that ES alone with filtering was effective enough (Kumar *et al.*, 2015).

ARIMA and ES both are good at predicting time-series datasets. These techniques will be evaluated against each other and the best will be chosen to use for the weather dataset.

2.2.4 Movie Recommendation based on various factors

Li *et al.*, (2016) made a movie recommendation system based on group-level sentiment analysis. The authors introduced KBridge, recommendation system consist of multiple data mining techniques and methods. The authors focused on user activities on

social media, which will affect the recommendation system based on previous behaviors. Another study was done by Li *et al.*, (2018), proposed a hybrid recommendation algorithm consist of the feature of movie and rating matrix of those movies. The higher the similarity matrix, the more inclination to recommend a movie to an end-user. The authors used to create a similarity matrix between users to make their decision more accurate and varied. Movie recommendation based on tags and ratings was also done in 2016. The authors used a social movie network (SMN), it contains a rating and tagging system for movies. Every user checks the movie ratings and tags before watching the movie on social media for critiques on the movie and then decide whether to go watch or not. The authors found that alone rating cannot recommend a movie and thus opted for the inclusion of social tags and ratings, which gives a better model in terms of accuracy (Wei *et al.*, 2016).

Thus, all the previous studied works in the movie recommendation domain use some types of ratings, social media, users' behaviors, and many more. There is no study done on recommending a movie based on the weather. The weather forecast will be done using supervised learning machine algorithms. This study will help movie streaming services and online websites to recommend a movie based on this parameter by making auto-recommendation of movies to the users based on the weather forecast of Dublin, Ireland. This can help the movie streaming industry to accurately recommend a movie to an end-user.

3. Methodology

In this proposed research, there are two models to be designed, one for forecasting the weather and second for classification of movies based on weather. CRISP-DM (Cross Industry Standard Process for Data Mining) methodology will be used for this research. CRISP-DM consists of six stages namely, Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. It is well known and robust methodology and used all over the industry (William Vorhies, 2016).

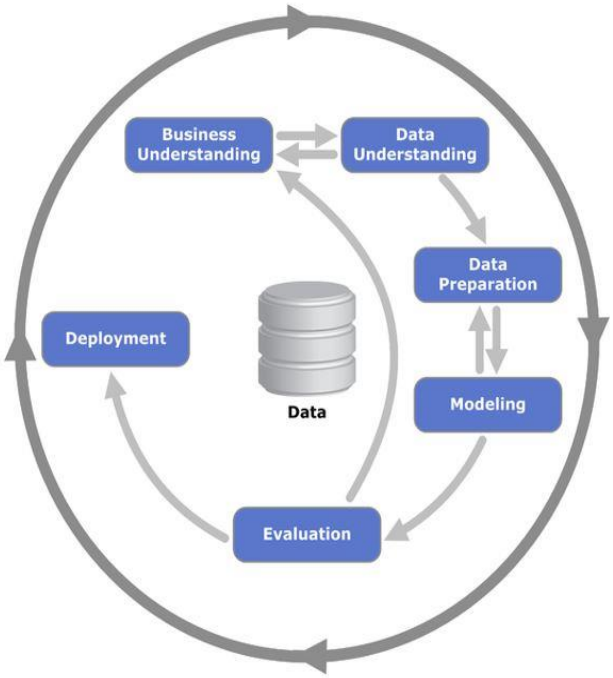

Fig. 1 Process Cycle of CRISP-DM

3.1 Architecture

The weather data of Dublin is used for prediction of weather model and train the model using machine learning algorithms for time series analysis like ARIMA and ES. It is forecasting the weather for the day. The prediction says the weather will be sunny, cold, or rainy. The dataset used for this model is sourced from Met Eireann website.

The prediction which then will be used for recommending similar types of tags of movies depending upon the weather like previously it did use case-based reasoning. Classification models are built in order to classify the movies into their respective tags.

There will be a couple of datasets used for this research, one from weather and second from movie datasets. Its process will be:

- Collecting data and making it to the required data type format.
- Predicting weather from the model and then classifying them into weather tags.
- Computing the accuracy and efficiency of the model.

The movie dataset is sourced from the MovieLens database and will be used to train and test the models. The flowchart of the architecture is represented below:

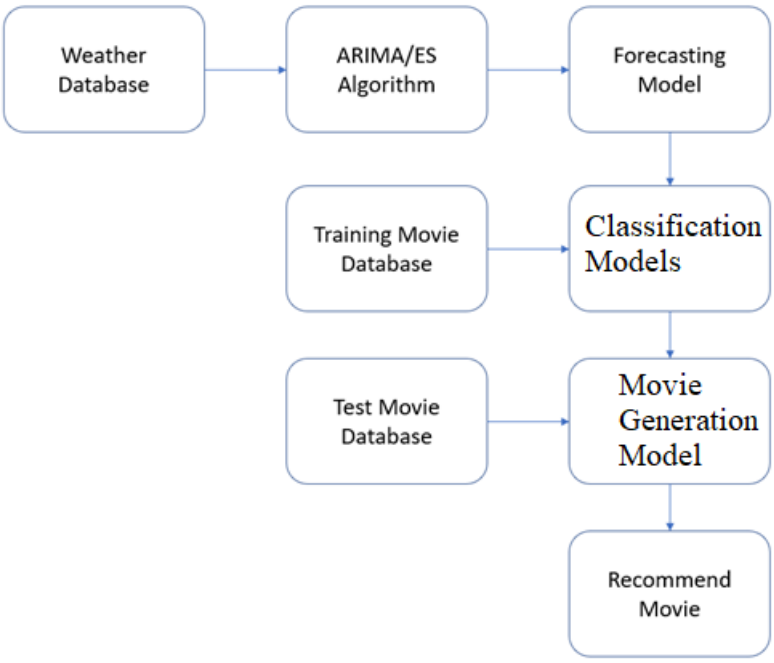

Fig. 2 Architectural Design of the Model

3.2 Data Gathering

The weather data is downloaded from the Met Eireann website (www.met.ie) which is publicly available. The datasets will have limited factors available if more factors needed the data can be requested from the organization under the act of Environment Regulations (AIE) (Eireann; 2008). The weather characteristics that are crucial for predicting the weather are air temperature, humidity, wind speed, the moisture of the air, and pressure along with the ground data. All this characteristic is considered as the independent factor which is relevant for the prediction of the type of weather. Attributes are shown in Table 1.

Sr. No.	Data Type	Variable	Meaning
1	string	date	Date
2	num	maxtp	Maximum Temperature
3	num	mintp	Minimum Temperature
4	num	rain	Precipitaion Amount
5	num	soil	Mean Soil Temperature
6	num	sun	Sunshine Duration (Hours)
7	num	evap	Evaporation
8	num	ind	Indicator
9	num	gmin	Grass minimum temperaure
10	num	pe	Potential Evotransipration

Table 1 Data Description of Weather Dataset

The Movie dataset is from MovieLens.org which is publicly available. A movie recommendation service. It contains 27753444 ratings and 1108997 tag applications across 58098 movies. These data were created by 283228 users between January 09, 1995, and September 26, 2018. This dataset was generated on September 26, 2018. No demographic information is included (Harper and Konstan, 2015).
The data are contained in the files `genome-scores.csv`, `genome-tags.csv`, `links.csv`, `movies.csv`, `ratings.csv` and `tags.csv`.
The data in the files are identified by different unique IDs used for allocating each movie. Each movie is allocated a tag defined by multiple users to correctly identify the type/theme of the movie. These tags vary from happy, sad, volcano, earthquake, sunny, rainy, so on depicting either mood, external conditions in which the movie is watched. The movie data containing tags related weather like snow, sunny, cloud, rain was extracted from all the tags respective to the movies using formulae in excel. Features are shown in Table 2.

Sr. No.	Variable	Data Type	Meaning
1	userId	num	Users ID
2	movieId	num	Movies ID
3	tag	string	Tags
4	title	string	Titles
5	genres	string	Genres
6	rating	num	Rating

Table 2 Data Description of Movie Dataset

3.3 Methodology

For this study, 2 models are developed and are explained below.

3.3.1 Weather forecast model

ARIMA is a type of time series analysis, a part of a machine learning algorithm that predicts the data. It is a statistical model. It allows the transformation of nonstationary series into stationery with the help of various parameters passed onto the model. Wilson, (2016) reviewed the Box and Jenkins method for time series analysis. RMSE is considered to be the performance measure of the models.

Exponential Smoothening is also a predicting technique (Sahay *et al.*, 2018). This uses the difference of last predicted value from the actual value and then gives the output predicted values. The formula for predicting the next values is:

$$\hat{x}_{t+1} = \hat{x}_t + \alpha\left(x_t - \hat{x}_t\right), \qquad (1)$$

Where α is the smoothing constant, and t+1 is the predicted value. This is useful for short time series analysis.

Both will be compared with respect to their performance and then the best model to fit for this study will be chosen.

3.3.2 Methodology of Movie Model

Movie recommendation will always be dependent on the weather conditions. For this, the previous output will be used to determine future recommendations. Case-based reasoning is a supervised learning machine learning algorithm. It is a data mining technique, that is used by many researchers for various studies (Boral, Chaturvedi and Naikan, 2019). KNN and Random Forest will be used as a classification algorithm for the classification of the movie into their respective fields, and then recommended accordingly. The training datasets will contain 80% of the sample and 20% for the testing dataset. Both the models KNN and Random Forest will be compared and then the best model will be chosen.

3.4 Performance Measure

3.4.1 Evaluation of Weather Model

ARIMA and ES both will be compared by computing Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). It is the most widely used measure to calculate the accuracy of the time series models. Both use the difference between the actual values and predicted values, whichever is the lowest, it is the best model.

3.4.2 Evaluation of Movie Model

Supervised algorithms like K-Nearest-Neighbour (KNN) and Random Forest (RF) will be used to check which model performs better. KNN algorithm will choose a specific K value, which will indicate the number of neighbors and the RF algorithm will choose mtry parameter to split the nodes. Confusion Matrix is used to see in-depth results and accuracy will be checked.

4. Implementation

4.1 Data Preparation

The weather dataset from Met Eireann had many attributes that were not required for this study. Those attributes were removed in Excel. The dataset also contained some NA or 0 values in the dataset, for this is. na () function from R in used to remove them and replace them. For ARIMA/ES model, the data is required in time series structure, hence ts() function is used to convert the object into time series in R. The ts() object was also cleaned in order to not have any outliers which will affect the model, it was removed by tsclean() function in R.

The movie datasets were distributed among 6 files and VLOOKUP function of Excel was used to bring the data altogether in one sheet. Only the tags containing weather-related tags were retained by filtering and sorting the dataset. All the attributes were converted into numeric datatype as knn only accepts numeric values. All these changes were made and then stored into a csv file for further process.

4.2 Models

4.2.1 Implementing Weather Model

This study compares both the methods of forecasting with the ARIMA model and the Exponential Smoothening model. These methods and functions are in packages called forecast, ggplot2, and tseries in R. These models are evaluated on the basis of RMSE and MAPE values, lower the value, greater the accuracy. For Exponential smoothening, ets() is used as it takes all the value into consideration and forecasts the values. It figures the best alpha, beta and gamma to choose from the options and gives out the best fit model. ARIMA model uses auto.arima () function in R which takes all the possible p,d,q values in ARIMA(p,d,q), and chooses the best fit model for the dataset.

Three models are built to predict precipitation amount, sunshine duration, and temperature. The results are explained in the evaluation section. After predicting these values, labels are assigned accordingly the predicted value. As per Met Eireann, if the sunshine duration is more than 10 hours then, the day is considered as sunny[1], else if it rains more than 12 mm than it is labeled as rainy day[2], or it is declared as cloudy.

4.2.2 Implementing of Movie model

After labeling the day, as one of the three weather types, the predicted weather label is matched with the appropriate movie tags from the dataset and then it is picked and stored into another data frame. The tags are used to generate the movie model results and are used in the KNN and RF models. All the attributes are changed to numerical datatype and then removed any NA and then normalized the dataset in order to have a good dataset. The dataset is then sampled into 80/20 ratio, 80% for training and 20% for testing. For these algorithms, caret () library is used to have flexibility. Both the models are compared and the best one is chosen.

5. Evaluation

5.1 ARIMA vs ES

Sunshine is used in both the calculations for the models to decide which model is better. Ets () is used for exponential smoothening modeling and auto.arima () is used ARIMA modeling.

Sr. No.	Performance Metrics	ES	ARIMA
1	RMSE	2.104675	0.650132
2	MAE	1.641685	0.515108
3	MPE	-3.724124	-5.52782
4	MAPE	17.12265	17.45367
5	ACF1	0.9559697	-0.007858

Table 3 ARIMA vs ES

As seen from above, RMSE of ARIMA is much lower than the RMSE of ES, hence for this dataset ARIMA is chosen for forecasting the weather.

[1] https://www.met.ie/climate/what-we-measure/sunshine
[2] https://www.met.ie/climate/what-we-measure/rainfall

The auto.arima () function in R has calculated all the best fit ARIMA model for temperature, rain, and sunshine, they are ARIMA (3,0,2)(2,0,1)[30], ARIMA (2,0,3) and ARIMA (2,0,2) respectively.

```
Series: deseasonal_cnt
ARIMA(2,0,3) with non-zero mean

Coefficients:
          ar1      ar2      ma1      ma2     ma3    mean
      1.5901  -0.6745  -0.4597  -0.0175  0.1541  1.0903
s.e.  0.0199   0.0177   0.0212   0.0138  0.0231  0.0296

sigma^2 estimated as 0.08889:  log likelihood=-1354.89
AIC=2723.78   AICc=2723.8   BIC=2771.24
```

Fig 3 ARIMA model for rain

```
Series: deseasonal_cnt
ARIMA(3,0,2)(2,0,1)[30] with non-zero mean

Coefficients:
         ar1     ar2      ar3     ma1     ma2     sar1     sar2    sma1     mean
      0.6494  0.8652  -0.5323  0.9225  0.0449  -0.3795  -0.0017  0.3621  13.4998
s.e.  0.0213  0.0208   0.0194  0.0253  0.0228   0.2630   0.0147  0.2631   0.5334

sigma^2 estimated as 0.1556:  log likelihood=-3174.65
AIC=6369.3   AICc=6369.33   BIC=6437.09
```

Fig 4 ARIMA model for temperature

5.2 Evaluation of weather model

ARIMA came out to be the best model among all, and hence chosen for the forecasting of the weather. Labels/tags will be generated if the weather is predicted as rainy, sunny, or cloud. Therefore, 3 models are built for the labeling purpose and prediction (Wan Ahmad and Ahmad, 2013).

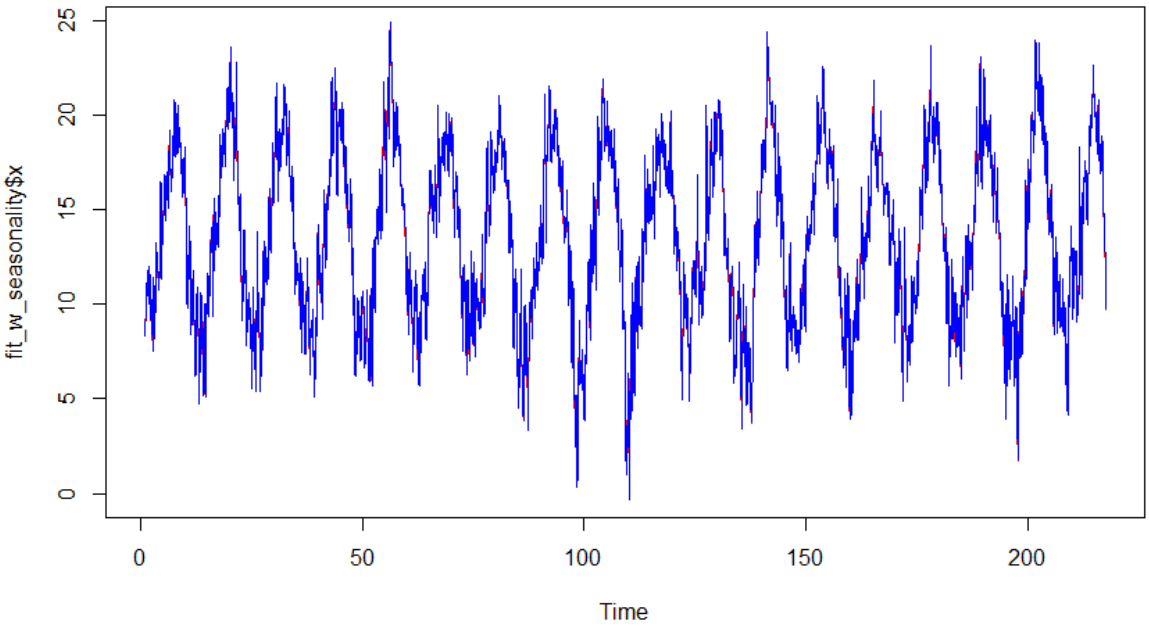

Fig 5 Model fitting to actual vs predicted the graph

In the above graph, red lines show the actual values and blue lines show the predicted values of the best-fit model. All the values of RMSE, MAE, MPE, MAPE, ACF1 of all the ARIMA models are depicted below in the table.

Sr. No.	Performance Metrics	ARIMA (3,0,2) (2,0,1) [30]	ARIMA (2,0,2)	ARIMA (2,0,3)
1	RMSE	0.39421	0.650132	0.298
2	MAE	0.310858	0.515108	0.2164
3	MPE	-0.810314	-5.52782	-55.85385
4	MAPE	3.4641	17.45367	118.4308
5	ACF1	-0.00143	-0.007858	0.02064

Table 4 Values of ARIMA models

5.3 Evaluation of Movie model

This model will classify the movies in 3 classes, namely cloudy, rainy, sunny. In this section, KNN algorithm and RF algorithm both are used to get the best fit model with the best accuracy amongst them. For KNN, caret () package is used for the algorithm. The author has used 10-fold cross validation and 3 repeats in order to get the best results out of the model. Value of K is decided by three methods,

KNN for k=19 (SQRT of Instances)	0.5280153
KNN for k=17 (Random)	0.5280153
KNN for k=21 (Best Value)	0.5362276

Table 5 KNN model evaluation

As from the above table, K=21 works best for this particular dataset giving 53.62% accuracy. Even Random Forest algorithm is used to build a model to check if it performs better than KNN.

RF mtry=2	0.5679101
RF mtry=3	0.5580296
RF mtry=4	0.56071

Table 6 RF model evaluation

In RF, mtry variable is chosen from the set of predictors, to form each split a different random set of variables. Mtry = 2 gives the best accuracy of almost 57%. RF accuracy is greater than then accuracy of KNN. Hence, RF is selected as the best-fit model for the dataset.

5.4 Discussion

This research classifies movies changing according to the weather. The weather prediction model which was chosen to be ARIMA forecasting model has RMSE of 0.298 which shows more accuracy than the ES model. Even the rainy model (RMSE 0.298) shows more accuracy than the temperature model (RMSE 0.394). As there was seasonality in the data, trends tend to help get more accuracy in the model. By the study of Wan Ahmed and authors, they concluded that ARIMA model is better than exponential smoothening model for forecasting a longer period of time (Wan Ahmad and Ahmad, 2013).

The comparison between K-nearest-neighbour model and Random Forest model shows that RF model is better than KNN for this dataset having mtry value of 2 with an accuracy of almost 57% which is justifiable with the sample of the dataset and classes available for the study. Thus, this proves that weather has an influence on movie suggestions and can be used to suggest a movie to users.

This research is not only limited to movie suggestion, but it can also be used in many fields like food takeout's/ delivery affected by weather conditions, productivity at the workplace, and many more. It can also be applied to other things like choosing a particular movie depending on the holiday season, behavioral pattern, and many more.

6. Conclusion and Future Work

This research will aid in the field of movies application to use weather data in order to suggest the movies with respect to the environment around the user. The movies that were tagged according to weather is used to classify them in the model. There are many things, weather data can be helpful like energy consumption, prevention of road accidents, traffic jams, and many more. The weather prediction model came out with the RMSE of 0.298 and the classification with an accuracy of 57% for this study.

We can use different classifiers to see if the accuracy can be increased by incorporating multiple algorithms. The limitation of this study is that it did not record the user's interaction and watching behavior, which would help to make the prediction more accurate. It can be included in future research.

As mentioned in discussions, the scope of this research can be prolonged in areas like food delivery business, or productivity at a workplace in the future. Similarly, this research can also be extended to see the feedback from the users of the current recommendations of movies.

Acknowledgment

Throughout the research study, I have received a great deal of support and assistance. I would first like to thank my supervisor, Dr. Cristina Muntean, whose expertise was invaluable in the formulating of the research topic and research proposal. I would like to acknowledge my colleagues from my master's at the National College of Ireland for their wonderful collaboration and suggestions.

References

About the Irish Film Industry | Screen Ireland (2018). Available at: https://www.screenireland.ie/about/about-the-irish-film-industry (Accessed: 12 August 2019).

Basu, S., Roy, S. and DasBit, S. (2019) 'A Post-Disaster Demand Forecasting System Using Principal Component Regression Analysis and Case-Based Reasoning Over Smartphone-Based DTN', *IEEE Transactions on Engineering Management*, 66(2), pp. 224–239. doi: 10.1109/TEM.2018.2794146.

Boral, S., Chaturvedi, S. K. and Naikan, V. N. A. (2019) 'A case-based reasoning system for fault detection and isolation: a case study on complex gearboxes', *Journal of Quality in Maintenance Engineering*, 25(2), pp. 213–235. doi: 10.1108/JQME-05-2018-0039.

Chavadi, C. A., Menon, S. R. and Sirothiya, M. (2019) 'Modelling the Effects of Brand Placements in Movies: An Investigative Study of Event Type and Placement Type', *Vision: The Journal of Business Perspective*, 23(1), pp. 31–43. doi: 10.1177/0972262918821227.
Cheng, W. *et al.* (2017) 'Predicting motorcycle crash injury severity using weather data and alternative Bayesian multivariate crash frequency models', *Accident Analysis & Prevention*, 108, pp. 172–180. doi: 10.1016/j.aap.2017.08.032.

Cook, Sam (2019) '60+ Netflix Statistics, Facts and Figures (2019 Version)', *Comparitech*, 14 March. Available at: https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/ (Accessed: 12 August 2019).

De Felice, M., Alessandri, A. and Ruti, P. M. (2013) 'Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models', *Electric Power Systems Research*, 104, pp. 71–79. doi: 10.1016/j.epsr.2013.06.004.
Department of Information Technology *et al.* (2019) 'Identification of advanced data analysis in marketing: a systematic literature review', *Journal of Economics and Management*, 35, pp. 18–39. doi: 10.22367/jem.2019.35.02.

Frankenfield, J. (2019) *How Data Analytics Work, Investopedia*. Available at: https://www.investopedia.com/terms/d/data-analytics.asp (Accessed: 12 August 2019).

Harper, F. M. and Konstan, J. A. (2015) 'The MovieLens Datasets: History and Context', *ACM Transactions on Interactive Intelligent Systems*, 5(4), pp. 1–19. doi: 10.1145/2827872.

J. Javid, R. and Jahanbakhsh Javid, R. (2018) 'A framework for travel time variability analysis using urban traffic incident data', *IATSS Research*, 42(1), pp. 30–38. doi: 10.1016/j.iatssr.2017.06.003.

Kalinić, H. *et al.* (2017) 'Predicting ocean surface currents using numerical weather prediction model and Kohonen neural network: a northern Adriatic study', *Neural Computing and Applications*, 28(S1), pp. 611–620. doi: 10.1007/s00521-016-2395-4.

Kamga, C. and Yazıcı, M. A. (2014) 'Temporal and weather related variation patterns of urban travel time: Considerations and caveats for value of travel time, value of variability, and mode choice studies', *Transportation Research Part C: Emerging Technologies*, 45, pp. 4–16. doi: 10.1016/j.trc.2014.02.020.

Kaminskas, M. and Ricci, F. (2012) 'Contextual music information retrieval and recommendation: State of the art and challenges', *Computer Science Review*, 6(2–3), pp. 89–119. doi: 10.1016/j.cosrev.2012.04.002.

Kumar, S. V. *et al.* (2015) 'INTEGRATION OF EXPONENTIAL SMOOTHING WITH STATE SPACE FORMULATION FOR BUS TRAVEL TIME AND ARRIVAL TIME PREDICTION', *TRANSPORT*, 32(4), pp. 358–367. doi: 10.3846/16484142.2015.1100676.

Li, H. *et al.* (2016) 'An intelligent movie recommendation system through group-level sentiment analysis in microblogs', *Neurocomputing*, 210, pp. 164–173. doi: 10.1016/j.neucom.2015.09.134.

Li, J. *et al.* (2018) 'Movie recommendation based on bridging movie feature and user interest', *Journal of Computational Science*, 26, pp. 128–134. doi: 10.1016/j.jocs.2018.03.009.

Malin, F., Norros, I. and Innamaa, S. (2019) 'Accident risk of road and weather conditions on different road types', *Accident Analysis & Prevention*, 122, pp. 181–188. doi: 10.1016/j.aap.2018.10.014.

Mitkov, A. *et al.* (2019) 'Forecasting the Energy Consumption in Afghanistan with the ARIMA Model', in *2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA). 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA)*, Varna, Bulgaria: IEEE, pp. 1–4. doi: 10.1109/ELMA.2019.8771680.

Moazami, A. *et al.* (2019) 'Impacts of future weather data typology on building energy performance – Investigating long-term patterns of climate change and extreme weather conditions', *Applied Energy*, 238, pp. 696–720. doi: 10.1016/j.apenergy.2019.01.085.

Nakatsu, R. and Edirisinghe, C. (2011) 'The Role of Movies and Telephony in the History of Communication Media', in *2011 Second International Conference on Culture and Computing. 2011 Second International Conference on Culture and Computing (Culture Computing)*, Kyoto, Japan: IEEE, pp. 69–73. doi: 10.1109/Culture-Computing.2011.21.

Pla, A. *et al.* (2014) 'Context-Aware Case-Based Reasoning', in Prasath, R., O'Reilly, P., and Kathirvalavakumar, T. (eds) *Mining Intelligence and Knowledge Exploration*. Cham: Springer International Publishing, pp. 229–238. doi: 10.1007/978-3-319-13817-6_23.

Sahay, R. *et al.* (2018) 'Exponential Smoothing based Approach for Detection of Blackhole Attacks in IoT', in *2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS). 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Indore, India: IEEE, pp. 1–6. doi: 10.1109/ANTS.2018.8710073.

Senior Psychometric Analyst — Mettl. Bangalore *et al.* (2019) 'Weather, Mood And Helping Behavior : A Comparative Study Between Two Indian Cities (Chennai And Mysore)', *JOURNAL OF PSYCHOSOCIAL RESEARCH*, 13(2), pp. 453–467. doi: 10.32381/JPR.2018.13.02.20.

Sinha, A. *et al.* (2019) 'Signaling effects and the role of culture: movies in international auxiliary channels', *European Journal of Marketing*, 53(10), pp. 2146–2172. doi: 10.1108/EJM-09-2017-0587.

Wan Ahmad, W. K. A. and Ahmad, S. (2013) 'Arima model and exponential smoothing method: A comparison', in. *PROCEEDINGS OF THE 20TH NATIONAL SYMPOSIUM ON MATHEMATICAL SCIENCES: Research in Mathematical Sciences: A Catalyst for Creativity and Innovation*, Palm Garden Hotel, Putrajaya, Malaysia, pp. 1312–1321. doi: 10.1063/1.4801282.

Wei, S. *et al.* (2016) 'A hybrid approach for movie recommendation via tags and ratings', *Electronic Commerce Research and Applications*, 18, pp. 83–94. doi: 10.1016/j.elerap.2016.01.003.

William Vorhies (2016) *CRISP-DM – a Standard Methodology to Ensure a Good Outcome*. Available at: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome (Accessed: 14 August 2019).

Wilson, G. T. (2016) 'Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1: TIME SERIES ANALYSIS: FORECASTING AND CONTROL,5TH EDITION, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Publis', *Journal of Time Series Analysis*, 37(5), pp. 709–711. doi: 10.1111/jtsa.12194.

Winoto, P. and Tang, T. Y. (2010) 'The role of user mood in movie recommendations', *Expert Systems with Applications*, 37(8), pp. 6086–6092. doi: 10.1016/j.eswa.2010.02.117.

Zhang, F. *et al.* (2019) 'Effect of introducing weather parameters on the accuracy of milk production forecast models', *Information Processing in Agriculture*, p. S221431731830355X. doi: 10.1016/j.inpa.2019.04.004.

Zhao, Z. *et al.* (2018) 'Improvement to the Prediction of Fuel Cost Distributions Using ARIMA Model', in *2018 IEEE Power & Energy Society General Meeting (PESGM)*. *2018 IEEE Power & Energy Society General Meeting (PESGM)*, Portland, OR: IEEE, pp. 1–5. doi: 10.1109/PESGM.2018.8585984.