Study of the effects of climatic conditions on alcohol consumption and the study of road accident fatalities

Soham Sameer More
School of Computing. National College of Ireland
(of Affiliation)
Dublin, Ireland
x19149240@student.ncirl.ie

***Abstract*— In this paper I study the effects of weather on the alcohol consumption pattern in the state of Iowa, USA. Also, the road accidents in the state of Iowa, USA are analyzed to study the pattern and factors affecting the fatality in an accident. Data Mining and Machine Learning methods like Multiple Linear Regression, Support Vector Regression, Logistic Regression, Naive Bayes and Support Vector Machine are used to conduct this study. It was found that the pattern of alcohol consumption is not related to the weather conditions in the state of Iowa, USA. The fatalities in a road accident were found to be dependent on the circumstances of the accident. The datasets in this study encompass the details of weather, alcohol consumption and road accidents in the state of Iowa, USA.**

Keywords*— *data mining, alcohol consumption pattern, climate, road accidents, fatalities

I. INTRODUCTION

A. Climatic effects on alcohol consumption

Alcohol is an integral part of many cultures and traditions across the world. Alcohol is consumed in social settings like weddings, funeral, festivals, etc. or in casual settings where friends and family get together to have a few drinks. Alcohol has many short term and long term effects on the body, especially the thermoregulatory response to cold. The scientific research has a varied opinion on the effects of alcohol consumption on the thermoregulation in cold weather [Alcohol ingestion and temperature regulation during cold exposure]. Haight and Keatinge [Failure of thermoregulation in the cold during hypoglycaemia induced by exercise and ethanol] studied the effects of alcohol (ethanol) consumption in a cold environment by ingesting a study groups with alcohol and increasing the alcohol concentration in successive studies. It was found that at a lower concentration of alcohol (0.34 g per kg body wt.) there is no change in the thermoregulatory mechanism of the body. However at a higher concentration of alcohol (0.79 g per kg body wt.) the thermoregulatory mechanism of the body is affected which reduces the core body temperature at a faster rate. In another study by Fox et al. [Effect of alcohol on thermal balance of man in cold water] it was observed that even though the alcohol concentration of the study group was relatively higher (0.86 g per kg body wt.), there was no impairment of thermoregulatory mechanism. This difference in the results can be attributed to the difference in the experimental methodology used in these studies.

Beker et al. [Human Physiology in Extreme Heat and Cold] also state that extreme cold stimulates the alpha-adrenoceptors which promotes heat conservation in the body. This is facilitated by cutaneous vascular constriction (narrowing of the blood vessels) and decreased peripheral profusion (decreased blood supply to peripheral organs) which helps in maintaining the core body temperature and blood flow to the vital organs. This results in decreased skin and muscle temperature.

Alcohol consumption leads to flushing of skin [Alcohol - the Body and Health Effects - Book]. This means that the blood vessels in the skin expand leading to increased blood flow to the skin. This increases the temperature of the skin which leads to (1) increased sweating, (2) peripheral vasodilation (dilation of blood vessels in the skin) and (3) reduced shivering [Alcohol ingestion and temperatl:1re regulation during cold exposure].

The objective of this study is to analyse if there is any correlation between cold weather and consumption of alcohol in the state of Iowa, USA. The sales of alcohol is considered as the dependent variable in this study as it is a direct indicator of the consumption of alcohol.

B. Fatalities in road accidents

Road safety and road accidents have been a concern since the advent of motorized vehicles in the early 20th century [Thinking about the history of road safety]. Accidents are caused due to human error as well as unfavorable environment conditions like weather, road condition, lighting conditions, junction type, etc. [Real-time Traffic Accident Severity Prediction using Data Mining Technologies]. Road accident injuries and fatalities are a concern for those involved in the accident as well as the first response teams and emergency medical services (EMS). This study aims at predicting the fatality in an accident based on the circumstances of the accident like the road surface type, light conditions, junction type, drug or alcohol influence, weather

type, etc. This prediction can prove to be vital for the EMS in saving lives.

Road accidents data in the state of Iowa, USA for the year 2015 and 2016 is used in this study.

II. Related Work

A. Climatic effects on alcohol consumption

B. Fatalities in road accidents

III. Data Mining Methodology

Both the studies were conducted in accordance to the cross-industry process for data mining (CRISP-DM) methodology as it provides a structured approach to a data mining project. Each step of the CRISP-DM is detailed below for both the studies.

A. Business understanding

1) Climatic effects on alcohol consumption

Objective - The objective of this project is to analyze the alcohol sales data and the weather data in the state of Iowa, USA to find out if there is any correlation amongst them. This information is useful to forecast the alcohol sales based on the weather forecast. Also, this information can prove to be an early warning system for medical services in case of sudden drop in temperature as alcohol consumption in extreme cold can cause hypothermia and prove to be fatal.

Tools and techniques - Statistical computing language R is used to perform data pre-processing and model building. Data mining and machine learning techniques like multiple linear regression and support vector regression are used to predict the alcohol consumption (sales).

2) Fatalities in road

Objective – The objective of this study is to find the relation between a fatality in a road accident and the circumstances of an accident like the weather, road conditions, cause of accident, type of impact, single car or multiple cars etc. This information and prediction can be important for emergency medical services and local police departments.

Tools and techniques - Statistical computing language R is used to perform data pre-processing and model building. Data mining and machine learning techniques like Logistic Regression, Naïve Bayes Classifier and Support Vector Machine are used to predict the alcohol consumption (sales).

B. Data understanding

1) Climatic effects on alcohol consumption

- Data collection -

i. The alcohol sales data was acquired from the Iowa state government website. This dataset provides data for every sale made to a dealer by the state of Iowa. Each sale entry consists of data fields like date and time, store name, store location, item purchased, category of alcohol, quantity, volume and price of alcohol sold. This dataset was in comma separated values (csv) format and had 4464376 rows and 24 columns of data out of which 99348 values for the county column were missing. As the county column was used to combine the climate and alcohol sales data it the missing values populated with values using a python script and FCC census API using the latitude and longitude data available in the dataset.

ii. National Centers for Environmental Information - This dataset gives location and station wise data of climate conditions like temperature, humidity, precipitation, snowfall, etc. Data is available from multiple stations spread over the state of Iowa, USA. This dataset has a lot of missing values as every station reports a particular type of data. This data was consolidated by date and county to get a concise dataset. This raw data was split over two csv files with 10773 and 106654 rows and 62 columns was downloaded from the National Centers for Environmental Information portal.

- Data properties -

i. The state of Iowa, USA is an alcohol beverage control state, which means that the state holds a monopoly on wholesaling alcohol in the state. The alcohol sales data contains the details of each, and every liquor sale made by the state of Iowa to liquor stores. It contains columns such as Invoice Number, Date, Store Number, Location and Address, County, Alcohol Category, Item Number and Description, Number of Bottles sold, Cost and Retail value of a bottle, Volume of Liquor sold and the Sale in US Dollars.

ii. The weather datasets follow the Global Historical Climatology Network (GHCN) - daily documentation format. This is a well-defined documentation method and is used by GHCN which is an integrated database of climate summaries from land surface stations across the globe. Every data column has an associated attributes column which contains information about the data collection time and quality of data. The column names are in GHCN abbreviation format. The documentation from GHCN was used to understand values in each column. Factors like minimum and maximum temperature, precipitation and snowfall were used in this study.

2) Fatalities in road accidents

- Data collection -

The road accidents data was acquired from the Iowa state government website. This dataset has a record of every accident in Iowa state. This dataset was in comma separated values (csv) format and had 1048575 rows and 37 columns. The details and circumstances of each and every accident are recorded in this dataset.

- Data properties -

The accidents dataset contains details like Department of Transportation (DOT) case number, Crash Date, Location information, Major Cause, Crash Manner, Crash Severity, Surface Conditions, Drug or Alcohol involvement, Light Conditions,

Weather, Roadway Junction, First Harmful Event, Number of Vehicles and Occupants etc.

C. Data preparation

1) Climatic effects on alcohol consumption

i. Cleaning the alcohol data

The date column was first converted to a standard format using the anytime package. The dataset was then filtered to use the data for the years 2015 and 2016. This data had 99348 missing county values. County values play an important part in this study as the alcohol and weather datasets are combined using Date and County information. These missing values were populated using the FCC census API using the latitude and longitude
information present in the data.

An aggregate dataset was created from this dataset. Sales of all the stores in a county for a date were added to give a single row which gives the total sales of liquor in a county by date.

ii. Cleaning the weather data

Weather dataset contained 56 columns with data readings regarding weather and a few more columns specifying the location, date and station name of the station where the reading was recorded. These 56 columns are 26 pairs (values, attributes) of readings. Most of these columns contained NA values and these columns were not used in this study. As a result, these columns were deleted from the dataset. Columns such as Weather Type, Daily percent of possible sunshine, Minimum soil temperature, Maximum soil temperature, Daily total sunshine, etc. were removed from this dataset as majority of these columns contained NA values. The columns that were used in the final study are Maximum temperature, Minimum temperature, Snowfall, Snow depth, Precipitation. Similar to the alcohol sales dataset, this dataset too was reduced to one row per date per county. This was achieved by taking the mean of all the readings for a county on any day. For example, if there are 10 readings for maximum temperature for county Dallas then the maximum temperature is taking as the mean of all these 10 readings.

iii. Combining the alcohol and weather data

Both the datasets had two common columns, namely Date and County. The datasets were combined using a condition similar to inner join used in relational database systems. The non-matching rows in both the datasets were ignored.

2) Fatalities in road accidents

The road accidents dataset had very less noise and missing values. The Occupants column had a few outliers that seemed like errors. These rows were not included in the study to reduce erroneous variance in the model.

D. Modeling

1) Climatic effects on alcohol consumption

Regression analysis was conducted in this study to predict the alcohol sales based on the weather conditions like temperature, snowfall and precipitation. The following regression models were implemented:

a) Multiple Linear Regression

Multiple linear regression is a supervised parametric model that tries to fit a straight line that explains the variance in the dependent variable based on the variance in the independent variables. The equation of the line is as follows in this study:

Yi (Alcohol Sales in USD) = B0 + B1 (precipitation) + B2(Snowfall) + B3(Snow depth) + B4(Minimum Temperature) + B5(Maximum Temperature) + B6 (wind speed) + e (error)

Ridge, Lasso and Elastic Net regularization techniques were also applied in this study to reduce overfitting and to add an additional penalty parameter that aims to minimize complexity or reduce the number of features used in the final model.

b) Support Vector Regression

In contrast to ordinary least squares (OLS), the objective function of support vector regression (SVR) is to minimize the coefficients — more specifically, the L2-norm (L2 norm is calculated as the square root of the sum of the squared vector values) of the coefficient vector — not the squared error.

SVR is a non-linear regression model whereas Multiple Linear Regression is a linear regression method. SVR was chosen for this study in order to check how a non-linear regression method performs in contrast to a linear regression method.

2) Fatalities in road accidents

This study was conducted to predict road accident fatalities based on the accident circumstances and the environment of the accident. This dataset is unbalanced as the number of accidents with fatalities is very small as compared to the number of accidents without fatalities.

a) Logistic Regression

The target variable is a binary variable, Fatality which takes a value of 1 in case of an accident related death and 0 if there was no death in the accident. Logistic regression predicts the probability of occurrence of an outcome. The probability threshold of 0.5 is usually used to separate the outcome. However, since predicting a fatality in an accident is more important than predicting a non-fatal accident, a probability threshold of 0.4 was used.

Also, to remedy the unbalanced dataset, weights were used while building the logistic regression model.

b) Naive Bayes

Naive Bayes classifier usually performs at par with other classifiers and it is a very robust model that can be used in most of the classification problems.

c) Support Vector Machine

SVM uses a error margin threshold to calculate the hyperplane what will divide the data into distinct classes.

Both Linear and Radial kernels were used to build SVM models. Both these models were evaluated to check which one performs better in this study.

IV. Evaluation

A. Climatic effects on alcohol

Multiple Regression with regularization techniques and cross validation was performed in this study. The weather parameters like temperature, wind, snow and precipitation do not show any significant relation to alcohol consumption in the state of Iowa, USA as can be seen from the evaluation metrics below.

Method	RMSE	R^2
Multiple Regression	52440.06	0.1153669
K-Fold Multiple Regression		0.1093796
Ridge Regression		0.1153834
Lasso Regression		0.1064434
Elastic Net Regression		0.10647
SVR	56511.16	0.06304156

This means that either the alcohol sales in not related to weather conditions or the weather conditions in this specific geographic region do not influence alcohol consumption.

B. Fatalities in road accidents

1) Logistic Regression

Accuracy = 0.97295006948917
Confusion Matrix

Actual Values	Predicted Values	
	0	***1***
0	28648	37
1	761	55

F1 : 0.9863
Recall (1) = 0.06740196

2) Logistic Regression with Weights

Accuracy 0.734381885359818
Confusion Matrix

Actual Values	Predicted Values	
	0	***1***
0	20981	7704
1	132	684

F1 : 0.84264
Recall (1) = 0.8382353

3) Naïve Bayes

Actual Values	Predicted Values	
	0	***1***
0	30272	140
1	72	822

F1 : 0.9965
Recall (1) = 0.9194631

4) Support Vector Machine

Linear Kernel

Actual Values	Predicted Values	
	0	***1***
0	30403	9
1	883	11

Recall (1) = 0.01230425

In the study of road accident fatalities, the correct detection of a fatality is more important than the correct detection of a non-fatal accident. To correctly evaluate the models for this criterion, I used the recall for detection of a fatality in the test dataset as a metric of evaluation. Three classification models were used in this study, Logistic Regression, Naïve Bayes and Support Vector Machine. From the evaluation metrics mentioned above we can see that the Naïve Bayes model outperforms rest of the models.
The recall for detection of a fatality in an accident is 91.94% in Naïve Bayes classifier. This means that the model detects a fatality in an accident correctly 91.94% of the time. Which seems like a reasonably well performing model.

References