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

Soham Sameer More
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x19149140@student.ncirl.ie

***Abstract*—This is a proposal for the study of how the climate affects the alcohol consumption patterns in the state of Iowa, USA and study of the circumstances in which road accidents occur and the possible relation with fatalities and injuries will also be undertaken for the state of Iowa, USA.**

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

I. Motivation

A. Climatic effects on alcohol consumption

Alcohol has been an integral part of cultures across the globe for many centuries. The consumption of alcohol occurs either during specific occasions like a wedding, a funeral, a birthday, etc. or in a casual setting where friends and family get together and have a couple of drinks. This study will attempt to find if, apart from the above-mentioned events do any climatic factors affect the drinking pattern in people? This study can prove to be a stepping stone for other alcohol consumption related studies such as *mishaps due to alcohol consumption*, *health-related issues due to alcohol consumption*, *the study of alcohol consumption and domestic violence*, etc. Data on climate and liquor sales in the state of Iowa, USA for the years 2015 and 2016 are used to conduct this study [1] [3].

B. Fatalities in road accidents

Road accidents are inevitable, however, the pain and suffering that follows due to injuries and unfortunate fatalities should be minimized - this is the motivation behind the study of this dataset. The goal of this study is to find out if and how the circumstances of an accident like weather, road surface, location, etc. are in any way related to major life-threatening injuries or fatalities in accidents. Data on road accidents in the state of Iowa, USA for the years 2015 and 2016 is used to conduct this study [4].

II. Research Question

- How are alcohol sales correlated to weather in the state of Iowa, USA?
- Can fatalities or major injuries be predicted by the circumstances of an accident?

III. Initial Review

A. Climatic effects on alcohol consumption

Ventura-Cots et al [5] conducted a study of the relation between colder climates specifically the mean temperature and the number of hours of sunshine and alcohol consumption. Countries were categorized into tropical, dry, temperate, continental and polar in their study and religious subgroups that affected the alcohol consumption were removed from the study to avoid bias. The results confirmed that in the U.S., the colder climate had a positive correlation with alcohol consumption i.e. the colder the climate higher the consumption of alcohol. Hagstr¨om et al [6] studied the effects of fewer sunlight hours and it's the correlation with increased alcohol consumption in Nordic Northern European countries. Their study found alcohol consumption had a negative correlation to the temperature and sunlight hours i.e. the colder the climate lesser the consumption of alcohol. This was in contrast to the Southern European countries where they saw a positive correlation between cold climate and alcohol consumption. This shows that the relation between climate and alcohol consumption may be dependent on the geographies. This study intends to find if the study by Ventura-Conts et al [5] about the US also holds true for its state Iowa.

B. Fatalities in road accidents

Liling Li et al [7] studied the road accident fatalities using data mining techniques and with a focus on providing better emergency medical services (EMS) to the victims of road accidents. Their study found that environmental factors like weather, road surface, and light conditions do not affect the fatality rate whereas other factors like driver being drunk or not, type of collision location of the collision have a strong association with fatalities. Pisano et al [8] studied the U.S. highway crashes in adverse weather conditions and found that the weather has a significant impact on road safety. Almost a quarter of road accidents in the US are weather-related. Their study concluded that more research is needed on weather-related crashes to help understand the factors affecting the crashes and to take preemptive measures to avoid fatalities.

IV. Data Sources

A. *Iowa State Liquor Sales Data*

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, and price of alcohol sold [1]. This dataset has 4464376 rows of data out of which 99348 values for the county column were missing. As the county column will be used to combine the climate and alcohol sales data it is important to handle the missing values correctly. These missing values were populated with data using a python script and FCC census API using the latitude and longitude data available in the dataset [2].

B. *Iowa State Climate Data*

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 [3]. This dataset has a lot of missing values as every station reports a particular type of data. This data will be consolidated by date and weather station to get a concise dataset. Missing values will be approximated for variables like temperature, rainfall, wind speed, etc. wherever possible.

C. *Iowa State Road Accidents Data*

This dataset provides historic data about road accidents in the state of Iowa, USA. Data fields include date and time of the accident, cause of the accident, location, property damage, fatalities, injuries, number of cars and passengers, etc. are provided in the dataset [4].This dataset contains details of 110472 road accidents spanning two years (2015 and 2016). This is a very clean dataset and does not have missing values for variables that will be used for this study.

V. Machine Learning Methods

R will be used to perform all the core operations whereas Python will be used for data cleaning and pre-processing in these studies. Model building and testing will be done in R using appropriate packages.

A. *Climatic effects on alcohol consumption*

This study will predict the sales of alcohol-based on climatic conditions. The following regression methods will be used in this study:

- Multiple Linear Regression - The simplest and perhaps the most effective method for regression analysis. This model was chosen as it has good interpretability and since the feature weights are generated during the model building process this method is very fast at predicting the target. This study will also try to avoid overfitting by using regularization methods of ridge and lasso regression.
- Polynomial Regression - Real-world data rarely has a linear relationship. Polynomial regression provides us with an advantage of building non-linear models which can accurately mimic the real-world relationships.

B. *Fatalities in road accidents*

This study will predict if there is a possibility of a fatality or major injury in an accident based on the circumstances in which the accident occurred. The target is boolean in nature. It will be true if there is a possibility of a fatality or a major injury in an accident and false otherwise. The following classification methods will be used in this study:

- Naive Bayes - This method provides a probabilistic approach to classification. This is fast and easy to implement but the biggest disadvantage is the requirement of predictors to be independent. In most real-life cases, the predictors are dependent, this hinders the performance of the classifier. However, since this method is known to perform at par with other methods this was chosen to study the accident data.
- Logistic Regression - In this method, the target variable takes binary values - yes or no. Since the target variable in this dataset is binary, this is an ideal method to study the dataset. It uses a linear model equation under the hood to provide binary categorical output.
- Support Vector Machine (SVM) - This method is highly effective in higher dimensions and the outliers have a negligible effect on the model. Since the injuries and fatalities in the datasets can have outliers, this model might produce better results as compared to other classification models.

VI. Evaluation Methods

The k-fold cross-validation method will be used on both the studies to estimate the hyperparameters of the models. The following metrics will be used to evaluate the models.

A. *Climatic effects on alcohol consumption*

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared evaluation methods will be used to analyze the regression models. RMSE and MAE are similar in many ways however since RMSE method squares the errors before taking a square root it is useful to identify large errors that are particularly undesirable [9]. R-squared is easier to interpret since the values are always between zero and one.

B. *Fatalities in road accidents*

Since the road accidents' dataset is imbalanced, evaluation methods like accuracy cannot be used to analyze the models [10]. Confusion Matrix will be used to evaluate the classification models, specifically the Precision-Recall parameters and the F1 statistic.

References

[1] ”Iowa Liquor Sales”, Data.iowa.gov, 2020. [Online]. Available: https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy. [Accessed: 25- Feb- 2020].

[2] ”FCC Area API”, Geo.fcc.gov. [Online]. Available: https://geo.fcc.gov/api/census/. [Accessed: 12- Mar- 2020].

[3] NCEI, ”Daily Summaries Location Details: Iowa, FIPS:19 — Climate Data Online (CDO) — National Climatic Data Center (NCDC)”, Ncdc.noaa.gov, 2020. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/FIPS:19/detail. [Accessed: 25- Feb-2020].

[4] ”Vehicle Accidents in Iowa by Location”, Data.iowa.gov, 2020. [Online]. Available: https://data.iowa.gov/Crashes/Vehicle-Accidents-in-Iowa-by-Location-Last-Ten-Yea/5xg3-s5yb. [Accessed: 25- Feb- 2020].

[5] Ventura-Cots M, Watts AE, Cruz-Lemini M, Shah ND, Ndugga N, McCann P, et al. Colder weather and fewer sunlight hours increase alcohol consumption and alcoholic cirrhosis worldwide. Hepatology (Baltimore, Md). 2018. Epub 2018/10/17. https://doi.org/10.1002/hep.30315 PMID: 30324707

[6] Hagström H, Widman L, von Seth E (2019) Association between temperature, sunlight hours and alcohol consumption. PLoS ONE 14(9):e0223312. https://doi.org/10.1371/journal.pone.0223312

[7] L. Li, S. Shrestha and G. Hu, ”Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques”, Central Michigan University, USA, 2017.

[8] Pisano, Paul Goodwin, Lynette Rossetti, Michael. (2008). U.S. highway crashes in adverse road weather conditions.

[9] J. Wesner, ”MAE and RMSE — Which Metric is Better?”, Medium, 2016. [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. [Accessed: 06- Mar- 2020].

[10] B. Rocca, ”Handling imbalanced datasets in machine learning”, Medium, 2019. [Online]. Available: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28. [Accessed: 07- Mar- 2020].