**Predictive Modeling for Assessing Drunk Driving Propensity: Analyzing Age, Gender, Alcohol, and Cigarette Usage Factors**

Anamarys Arroyo-Lloret

Department of Public Health, Temple University

HIM 5102: Applications of Computer Programming in Health Informatics

Dr. Jay Patel

December 10, 2023

**Abstract**

Drunk driving remains a critical issue in contemporary society, prompting extensive research and analysis over the years. Previous studies have highlighted varied outcomes regarding the impact of drunk driving on men and women, revealing a notably higher prevalence among men. This study draws on data obtained from the 1996 National Survey on Drug Use and Health (NSDUH). Utilizing this dataset, we developed a predictive model to evaluate the likelihood of individuals driving under the influence, considering factors such as gender, alcohol intake, weight, employment status, and cigarette usage.

*Keywords:* Drunk driving, gender differences, predictive modeling, NSDUH, risk assessment

**Introduction**

Driving under the influence (DUI), also known as impaired driving, involves operating a motor vehicle after consuming alcohol (Addiction Center, n.d.). In 2020, the National Highway Traffic Safety Administration (NHTSA) reported 11,654 fatalities from alcohol-impaired driving incidents, accounting for 30% of total motor vehicle traffic fatalities in the United States. Annually, an estimated 1.5 million individuals are arrested for driving under the influence of alcohol, highlighting the persistent severity of impaired driving as a critical issue in traffic safety and public health nationwide (Addiction Center, n.d.).

Previous research indicates that alcohol significantly impairs both men and women, with women exhibiting greater impairment than men (Miller et al., 2009). Moreover, men demonstrate higher rates of current drinking and heavy drinking compared to women (Wilsnack et al., 2009). Schwartz & Beltz (2018) observed higher rates of male intoxicated drivers; nevertheless, there is a concerning increase in the rates of female intoxicated drivers.

Utilizing data collected from the National Survey on Drug Use and Health (NSDUH) from 1996, we aim to construct a predictive model assessing the probability of individuals driving under the influence, considering variables like gender, alcohol consumption, weight, work status, and cigarette use.

**Methods**

**Dataset Preparation**

We began our data preparation process by accessing the National Survey on Drug Use and Health 1996 dataset. In R, we retrieved the dataset and proceeded to identify our variables of interest. Utilizing R, we crafted a subset tailored to our specific criteria and downloaded the refined dataset in an Excel format for further analysis. Through R-based computations, we established a new sample size (N=18,269). The variables of interest included demographic factors such as age, sex, race, and education, along with behavioral aspects including alcohol and cigarette usage patterns, work status, height, weight, overall health, and various alcohol-related metrics such as frequency of use and initiation age. The focus of our analysis rested on the dependent variable, specifically the number of days an individual drove under the influence in the past 30 days.

***Analytic Tools***

R served as the primary tool for data file conversions, cleaning, manipulation, and descriptive analysis throughout this study. Python was utilized specifically for the creation of the prediction model.

***Python Code***

Step 1: Import all necessary libraries

import pandas as pd

from sklearn.pipeline import Pipeline

from sklearn.impute import SimpleImputer

from sklearn.feature\_selection import SelectKBest

from sklearn.feature\_selection import chi2

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn import metrics

import matplotlib.pyplot as plt

Step 2: Read the new dataset previously created in R

drug\_data = pd.read\_excel('/Users/anabanana/Desktop/EPBI\_8208/FinalProject/Data/DrugData\_Subset.xlsx')print(drug\_data)

Step 3: Correlation Calculation

Code 3 – Calculate correlations among the variables to determine relationships between these variables and provide insight into potential predictors.

drug\_data.corr()

Step 4: Data Preprocessing and Feature Selection

Code 4 - To start, missing values in the 'VAL30USE' column were addressed and a threshold was set. Feature selection occurred using a pipeline integrating imputation and chi-squared tests, showing the top three features for the model. The dataset was divided for training and testing, enabling the creation of a logistic regression model after handling missing data. The model underwent thorough evaluation using multiple metrics, culminating in the visualization of its predictive performance via an ROC curve and AUC score.

drug\_data['VAL30USE'].fillna(0, inplace=True)

threshold = 1

drug\_data['VAL30USE'] = drug\_data['VAL30USE'].apply(lambda x: 1 if x >= threshold else 0)

X = drug\_data.iloc[:, 1:14]

Y = drug\_data['VAL30USE']

pipeline = Pipeline([

('imputer', SimpleImputer(strategy='mean')),

('selector', SelectKBest(score\_func=chi2, k=3))

])

pipeline.fit(X, Y)

selected\_features = pipeline.named\_steps['selector'].get\_support(indices=True)

print(f"Selected feature indices: {selected\_features}")

df\_scores = pd.DataFrame(pipeline.named\_steps['selector'].scores\_)

df\_columns = pd.DataFrame(X.columns)

features\_scores = pd.concat([df\_columns, df\_scores], axis=1)

features\_scores.columns = ['Features', 'Score']

features\_scores = features\_scores.sort\_values(by='Score', ascending=False)

print(features\_scores)

X = drug\_data[['IRALCFQ', 'POUNDS', 'WORKSTAT']] # top 3 features

Y = drug\_data['VAL30USE'] # target output

Y = Y.values.ravel()

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.4, random\_state=100)

feature\_imputer = SimpleImputer(strategy='mean') # Use mean imputation for missing values in the features

X\_imputed = feature\_imputer.fit\_transform(X)

X\_imputed = pd.DataFrame(X\_imputed, columns=X.columns)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_imputed, Y, test\_size=0.4, random\_state=100)

logreg = LogisticRegression()

logreg.fit(X\_train, y\_train)

y\_pred = logreg.predict(X\_test)

print(X\_test) # Test dataset

print(y\_pred) # Predicted values

print('Accuracy:', metrics.accuracy\_score(y\_test, y\_pred))

print('Recall:', metrics.recall\_score(y\_test, y\_pred, zero\_division=1))

print('Precision:', metrics.precision\_score(y\_test, y\_pred, zero\_division=1))

print('CL Report:')

print(metrics.classification\_report(y\_test, y\_pred, zero\_division=1))

y\_pred\_proba = logreg.predict\_proba(X\_test)[:, 1]

false\_positive\_rate, true\_positive\_rate, \_ = metrics.roc\_curve(y\_test, y\_pred\_proba)

auc = metrics.roc\_auc\_score(y\_test, y\_pred\_proba)

plt.plot(false\_positive\_rate, true\_positive\_rate, label="AUC=" + str(auc))

plt.title('ROC Curve')

plt.ylabel('True Positive Rate')

plt.xlabel('False Positive Rate')

plt.legend(loc=4)

plt.show()

**Results**

***Descriptive Statistics***

We conducted a descriptive analysis to compare individuals based on their alcohol consumption. Segmentation into 'alcuser' (alcohol users) and 'nonalcuser' (non-alcohol users) subsets allowed focused examination. Using R, we determined 7,700 alcohol users and 10,569 non-alcohol users. Utilizing 'summary' for an overview and 'table' for categorical variables (sex, race, cigarette, and alcohol use), we identified distinct patterns between groups, highlighting demographic and behavioral differences.

**Table 1. Comparison of Demographic and Behavioral Characteristics Between Alcohol Users and Non-Alcohol Users**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Alcohol User**  **(N=7700)** | **Non-Alcohol User**  **(N=10569)** | **Overall**  **(N=18269)** |
| **Sex** |  |  |  |
| Male | 3770 (49%) | 4004 (38%) | 7774 (43%) |
| Female | 3930 (51%) | 6565 (62%) | 10495 (57%) |
| **Race** |  |  |  |
| American Indian | 124 (2%) | 174 (2%) | 298 (2%) |
| Asian or Pacific Islander | 153 (2%) | 346 (3%) | 499 (3%) |
| Black | 1697 (22%) | 2927 (28%) | 4624 (25%) |
| White | 5726 (74%) | 7122 (67%) | 12848 (70%) |
| **Cigarettes – Ever Used** |  |  |  |
| Never Used | 1374 (18%) | 5901 (56%) | 7275 (40%) |
| Ever Used | 6326 (82%) | 4668 (44%) | 10994 (60%) |
| **Alcohol – Ever Used** |  |  |  |
| Never Used | 1 (<1%) | 5082 (48%) | 5083 (28%) |
| Ever Used | 7700 (100%) | 5487 (52%) | 13187 (72%) |

***Predictive Model***

The analysis utilized a dataset with 17 columns, exploring various factors potentially associated with instances of drunk driving. Through a rigorous selection process, three key predictors—'IRALCFQ,' 'POUNDS,' and 'WORKSTAT'—were identified, ranking highest in their predictive capability.

**Table 2. Classification Metrics for Driving Behavior Prediction**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Did Not Drive Drunk** | 0.90 | 0.98 | 0.94 | 6549 |
| **Did Drive Drunk** | 0.38 | 0.11 | 0.17 | 759 |

The model constructed based on these features demonstrated an accuracy of 78.1%, highlighting its ability to predict drunk driving behavior. Further evaluation metrics revealed a recall rate of 88.1%, indicating the model's capacity to accurately identify 88.1% of actual drunk driving cases. However, precision stood at 30.7%, revealing that while the model correctly classified a proportion of drunk driving instances, it also had false positive predictions. This model's comprehensive performance, characterized by an accuracy of 78.1%, suggests its potential utility in identifying cases of drunk driving based on the specified features, albeit with some room for improvement in precision.

**Discussion**

The prediction model displayed creditable performance metrics, with precision, recall, and F1-score providing insights into its predictive capabilities. The precision of 0.90 signifies the model's accuracy in correctly identifying individuals who did not engage in drunk driving, while the recall of 0.98 highlights its adeptness in capturing most actual instances where individuals refrained from driving under the influence. The F1-score (0.94) represents a balanced measure of precision and recall, confirming the model's overall competency in distinguishing non-drunk driving cases. Although the model exhibited a notable accuracy of 78.1%, effectively predicting instances of drunk driving based on the specified features, it also showed limitations. The precision score of 0.38 revealed false positives, indicating areas for improvement in identifying cases of drunk driving.

Using newer data and including more factors that affect drunk driving could improve how well the model predicts, making it more reliable and accurate. This would help tackle the issues we found with how the model performed and the mistakes it made in predicting drunk driving cases.

**Conclusions and Future Work**

In summary, this research has delved into the prevalent issue of drunk driving, utilizing the 1996 National Survey on Drug Use and Health (NSDUH) dataset to develop a predictive model. The model, based on factors such as gender, alcohol intake, weight, employment status, and cigarette usage, demonstrated an accuracy of 78.1% in predicting instances of drunk driving. While the precision, recall, and F1-score metrics provided valuable insights into the model's performance, the study also identified limitations, particularly in addressing false positives.

To enhance the model's predictive capabilities, future work should focus on incorporating newer data and considering additional factors that influence drunk driving behavior. A more comprehensive analysis involving a broader demographic range, spanning both genders and ages 21 to 65, would contribute to a more nuanced understanding. Additionally, attention to ethical considerations is essential, given the potential consequences associated with predicting behavior.

Furthermore, exploring in-depth personal questions and gathering a more diverse sample size could refine the model's accuracy and reliability. This approach would contribute to addressing the identified issues in the model's performance and improve its effectiveness in predicting drunk driving cases. Ultimately, this research lays the foundation for future endeavors to advance predictive modeling in the context of tackling the complex and critical issue of drunk driving.

**References**

Addiction Center. (n.d.). *Drunk Driving.*

https://www.addictioncenter.com/alcohol/drunk-driving/

Miller, M. A., Weafer, J., & Fillmore, M. T. (2009). Gender Differences in Alcohol Impairment

of Simulated Driving Performance and Driving-Related Skills. *Alcohol & Alcoholism.*

*44*(6), 586-593. doi: 10.1093/alcalc/agp051

National Highway Traffic Safety Administration (NHTSA). (n.d.). *Drunk driving.*

<https://www.nhtsa.gov/risky-driving/drunk-driving>

Schwartz, J., & Beltz, L. (2018).Trends in female and male drunken driving prevalence over

thirty years: Triangulating diverse sources of evidence (1985–2015).

*Addictive Behaviors. 84*, 7-12. <https://doi.org/10.1016/j.addbeh.2018.03.024>

Wilsnack, R. W., Wilsnack, S. C., Kristjanson, A. F., Vogeltanz-Holm, N. D., & Gmel, G. (2009).

Gender and alcohol consumption: patterns from the multinational genacis project.

*Addiction. 104*(9), 1487-1500. doi:10.1111/j.1360-0443.2009.02696.x.