Predicting People’s Alcoholic Status

By: Nathan Kim, Abel Lula, Jerry Shi, Jack Stapholz, Rachel Stokol Lec 1 Group 10

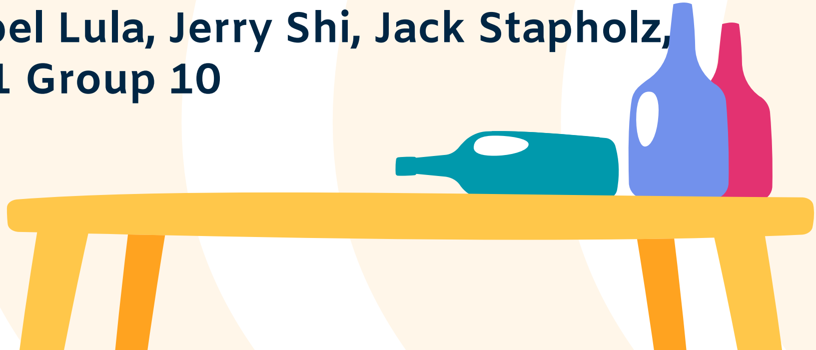


Table of contents

01
Introduction & Methodology
Project Description, Topic Discussion, & Data Setup

02
Model Selection
Model Creation Process

03
Results
Discussion of Results and Model Description

04
Conclusions
Final Discussion

01 Introduction & Methodology

Project Description, Topic Discussion, & Data Setup

Alcoholic Status Prediction

The data discusses a person's alcoholic status based off vitals such as:

- Height
- Weight
- LDL_Chole

The goal for this project is to utilize the predict() function in R and predict the testing data response variable, given the training data, with a modeling method or a combination of methods that we learned in class to predict future alcoholic statuses.

Testing and Training Dataset

Observations

Training Data: 70,000 Observations

Testing Data: 30,000 Observations

Dimensions

Training Data: 70,000 x 27

Testing Data: 30,000 x 26

Process

Clean Up Data

Address missing values and set up categorical variables as factors

Fit Models

Create multiple models of varying type and number of predictors

Test & Compare Models

Fit models, identify their accuracy rates, and select the best model according to accuracy

Identifying Missing Values

86% of observations in the complete dataset had at least one missing value

All 26 predictors had missing values with frequencies between **6-12%**

We employed three imputation methods to replace the missing values:

1. Mean, median, and mode
2. MICE
3. Hmisc

Imputing Missing Values

Average

- Impute NAs in numerical predictors with the mean or median
- Impute NAs in categorical predictors with the mode

MICE

- Use mice package in R to create a data frame of imputed values
- Merge imputed values and original data to replace NAs

Hmisc

- Use aregImpute() from Hmisc package in R to impute values (n.impute = 5)
- Merge imputed values from the 5th iteration of aregImpute() with original data to replace NAs

02

Model Selection

Model Creation Process

Models Tested

Models tested include: logistic regression, random forest, LDA, QDA, GLM CV, KNN, XGBoost, and GAM.

Logistic Regression

- GLM with 6 predictors: sex, age, weight, height, hemoglobin, smoking status
- Kaggle accuracy: 0.70513

Random Forest

- Random forest model with all predictors, 1000 trees, and 7 variables per tree
- Kaggle accuracy: 0.72906

GAM

- GAM with degree 6
- Kaggle accuracy: 0.73013

Predictor Selection

Step 1

Run backwards BIC stepwise regression to reduce predictors.

Step 2

Fit Generalized Additive Models to training data with degrees 3 & above.

Step 3

Run ANOVA between different degree GAMs with reduced & all predictors and analyze accuracy predicting training response variable.

Final Model

GAM model with all predictors and degree 6 natural splines on numerical predictors.

Pros and Cons of GAMs

Interpretability

Easy to interpret

Flexibility

Flexible functions can learn the distributions of predictors

Accuracy

Generalized Additive Models beat out many other methods with this data set

Variability

The increase in flexibility of a GAM model also increases variability (especially at extremes)

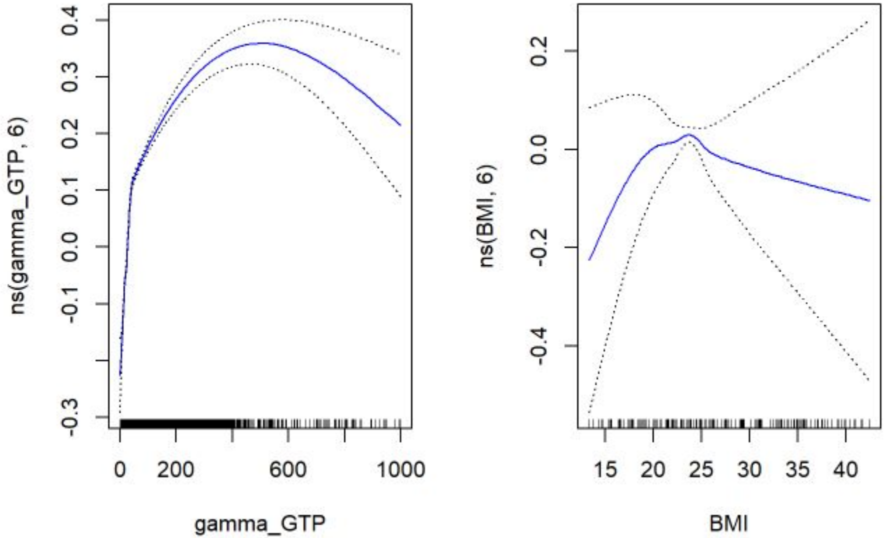
Numerical

Splines only work with numerical predictors, categorical variables remain unchanged

Choice

There are a number of different degrees and spline functions to choose from

Predictor Scatterplots



Generalized Additive Models learn the non-linear distributions of the numerical predictors

20 predictors

Number of numerical predictors

19 predictors

Number of statistically significant predictors

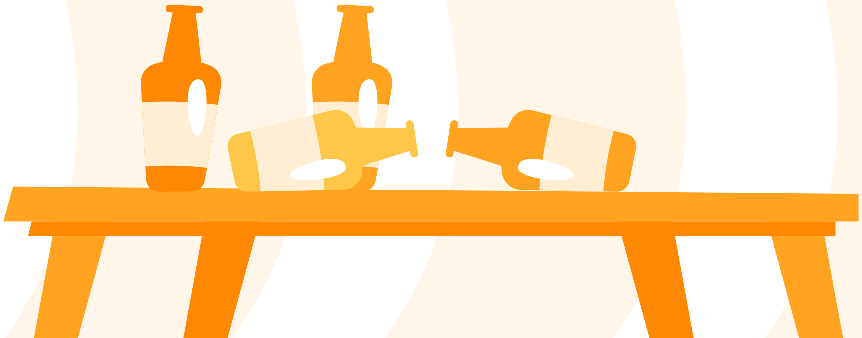
70,000

Total observations in training data

ANOVA of GAM with Reduced & All Predictors

Model	Resid DF	Resid Dev	DF	Deviance	P(>Chi)
Red	69927	12649			
Full	69873	12589	54	59.441	<2.2 e-16

ANOVA favors the GAM with all predictors with degree 6 over over a reduced degree 6 GAM



ANOVA of Degree 3 vs 6

ANOVA favors the higher degree GAM with degree 6 over the GAM with degree 3 splines

Model	Resid DF	Resid Dev	DF	Deviance	P(>Chi)
Deg 3	69930	12702			
Deg 6	69873	12589	57	112.69	<2.2e-16

Accuracy Predicting Response in Training Data

72.74%

GAM degree 3

GAM model with all predictors

72.77%

GAM degree 6

GAM model with predictors reduced by stepwise regression (BIC)

72.97%

GAM degree 6

GAM model with all predictors

Confusion Matrix of Predictions

Accuracy

Accuracy of predictions compared to training data

72.97%

Sensitivity

Sensitivity of the model

73.23 %

Training Set

Predicted		Yes	No
	Yes	25,364	9,399
	No	9,523	25,714

03

Results

Discussion of Results and Model Description

Final Model

GAM (with Degree 6)

Observations

30,000 Alcoholic Statuses
30x1

Predictors

26 Predictors

MCR/Rank

MCR: 0.26987
Rank: 30th

Important Predictors

Characteristics

Sex, age, height, weight, wasteline

Blood-Related

SBP, DBP, BLDS, tote_chole, HDL_chole, LDL_chole, triglyceride, hemoglobin, serum_creatinine, SGOT_ALT, SGOT_AST, gamma_GTP, BMI

Smoking

Smoking Status

Most Important Predictors

Predictors: Sex, Smoking.Status, Hemoglobin, gamma_GTP, height, age, triglyceride, HDL_chole, and waistline

From our model, we found that these predictors were most essential at predicting the alcoholic status. These predictors contributed the most to accurately predicting a person's alcoholic status.

04 Conclusion

Discussion/Limitations

Conclusion

Our most successful model was a GAM model using Hmisc imputations that had **73.01**% accuracy in predicting the alcoholic status for the test data.

We believe that if additional improvements were made to the imputation method for missing values, this model would be stronger and produce a higher accuracy rate.

Limitations

Model Complexity

- High Degree of 6
- Use of ALL Predictors
- Prone to Overfitting

Data (NA and Features)

- Missing Values in Data
- Lack of Health Domain Knowledge

GAM Model Assumptions

- Linearity
- Independence
- Smoothness
- Homoscedasticity

Better Data, Better Model

Using Original Full Dataset

- Same Random Forest model on 70,000 vs 1.6 mil rows
- 72% Accuracy vs 89% Accuracy
- Data and Features are as Important as Model Selection

Future Work

Suggestions

- Regularization Techniques (PCA, L1/L2, CV) -> Simpler Model
- Feature Engineering with Domain Knowledge
- Better Imputation Methods: KNN
- Using Ensemble or Combination of Methods

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**