Insurance Applicant Risk Analysis

Predicting applicant smoking and drinking habits based on health markers

GROUP – 11
Jagruta Advani
Marcus Martinez
Varsha Manju Jayakumar
Pratyush Rohilla
Soham S Bidyadhar

Problem Statement

Underwriting: Insurance firms assess risk to determine premiums and policy terms.

Health Assessment: Evaluates medical history, lifestyle choices (e.g., smoking, drinking), and current health to identify potential risks.

Predictive Need: Applicants might conceal smoking or drinking habits, so firms need predictive tools to accurately assess these risks.

Dataset

Data Source

Data is from South Korea's National Health Insurance Service via Kaggle

Variables

Includes standard health metrics and indicators like blood pressure, cholesterol, liver, and kidney function

Goal

Goal is to predict drinking and smoking habits and identify key indicators using a subset of 100,000 rows.

Agenda

Exploratory Data Analysis

Pre- Feature Engineering and Post-Feature Engineering Analysis

Feature Engineering

Features created to improve the accuracy of our prediction

Models Used

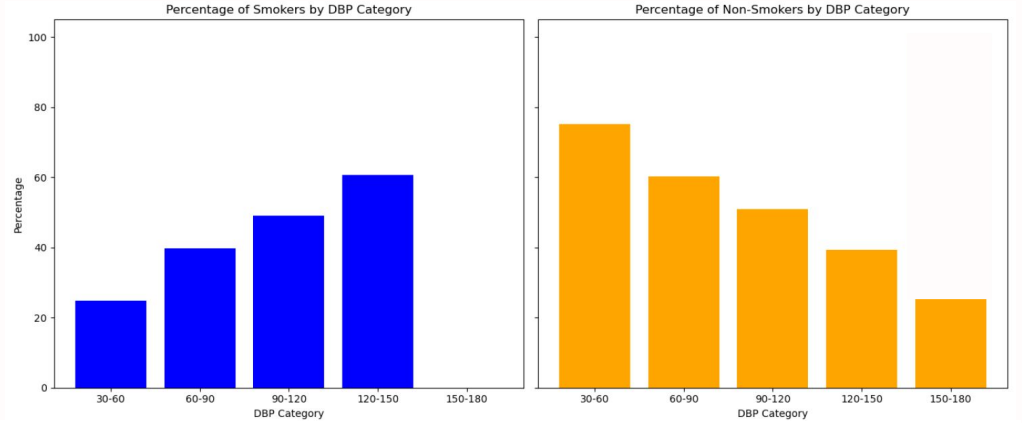
Classification algorithms that we chose for our predictions

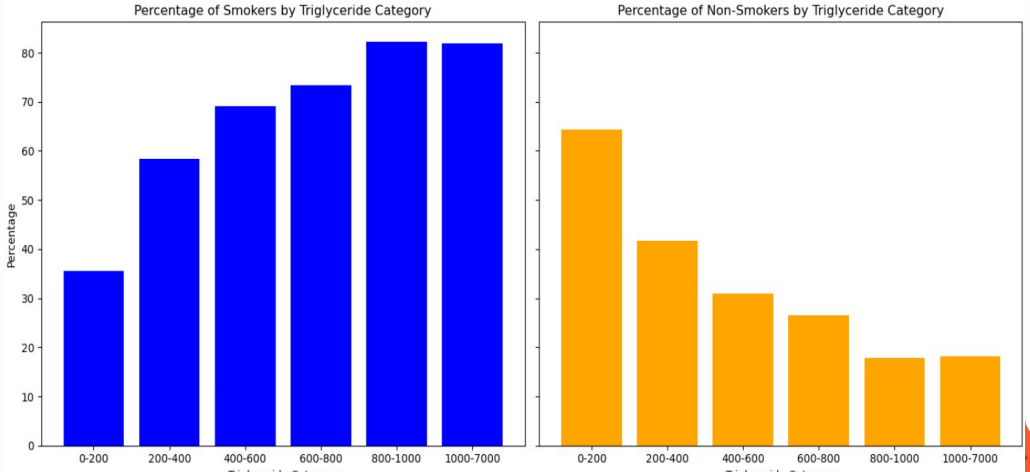
Insights and Conclusions

Best-fit model for our Dataset and conclusions of the analysis

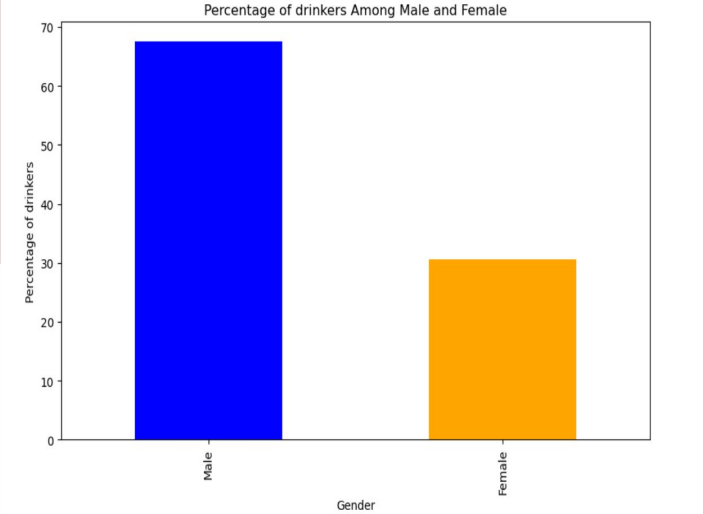
Preliminary Data Analysis and Visualizations

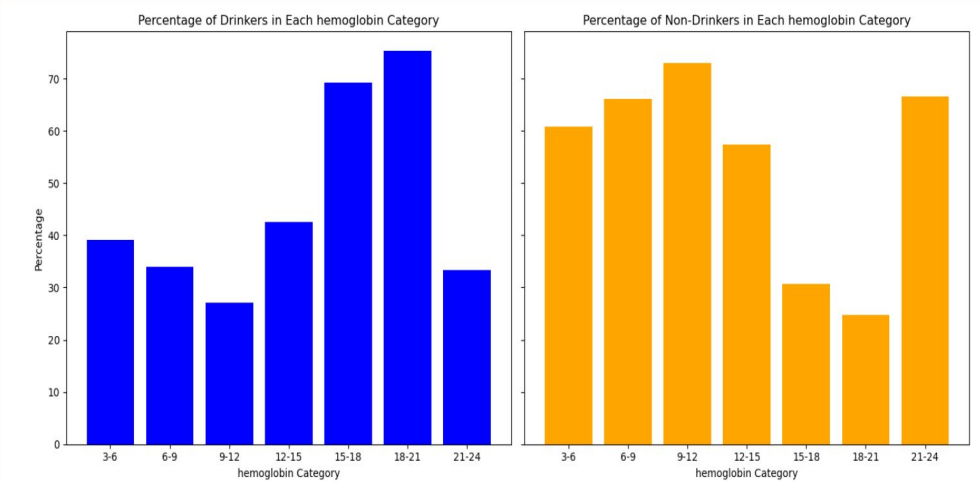
SMOKING

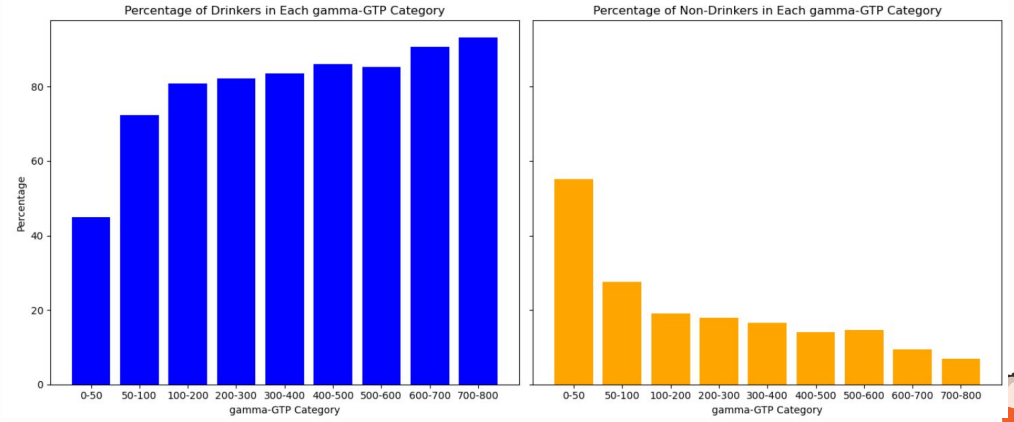




DRINKING


Percentage of drinkers Among Male and Female
Percentage of drinkers
0
10
20
30
40
50
60
70
Male
Female
Gender


Percentage of Drinkers in Each hemoglobin Category
Percentage of Non-Drinkers in Each hemoglobin Category
Percentage
0
10
20
30
40
50
60
70
3-6
6-9
9-12
12-15
15-18
18-21
21-24
hemoglobin Category

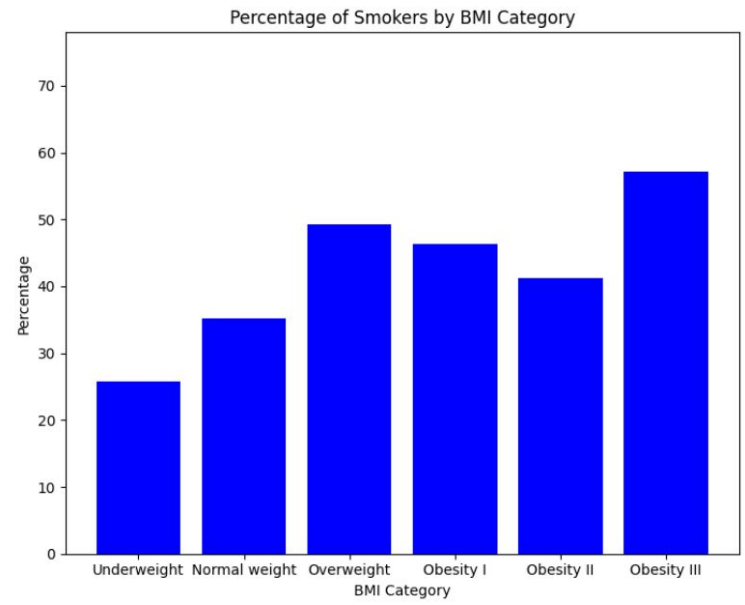

Percentage of Drinkers in Each gamma-GTP Category
Percentage of Non-Drinkers in Each gamma-GTP Category
Percentage
0
20
40
60
80
0-50
50-100
100-200
200-300
300-400
400-500
500-600
600-700
700-800
gamma-GTP Category

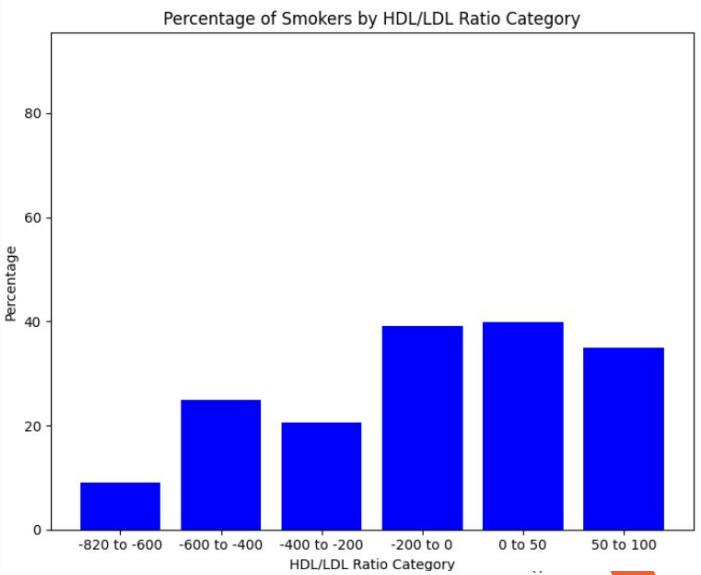
Feature Engineering

Key Variables	
Hearing_State	Classifies as faulty if either left/right hearing is faulty
HDL_LDL_Ratio	Ratio of HDL and LDL Cholesterol
BMI	(weight/height^2)
Total_HDL_Ratio	Ratio of Total Cholesterol and HDL
Liver_Enzyme_Ratio	SGOT_AST/ SGOT_ALT
Liver_damage_score	gamma_GTP + SGOT_ALT
Smoking_Status	Factorised smoking status - 0/1
Drinking_Status	Factorised drinking status - 0/1

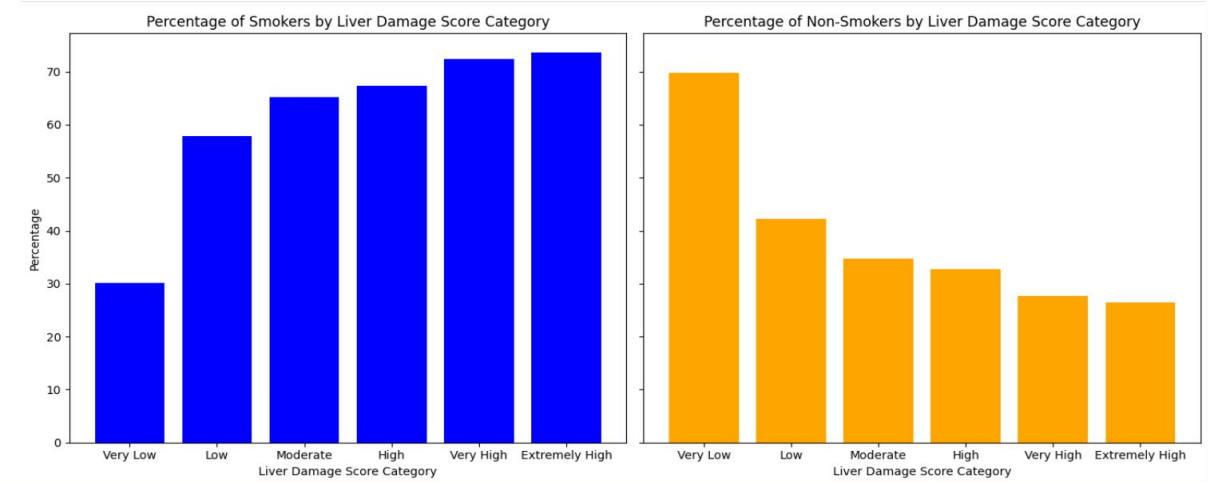
Post-Feature Engineering Analysis

SMOKING

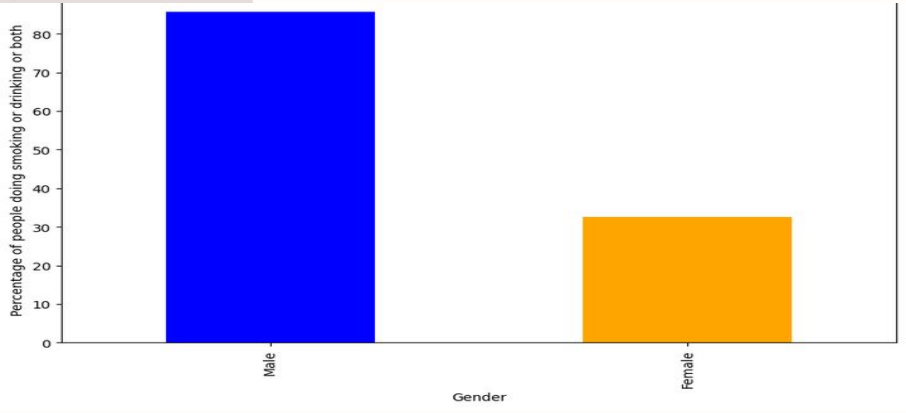


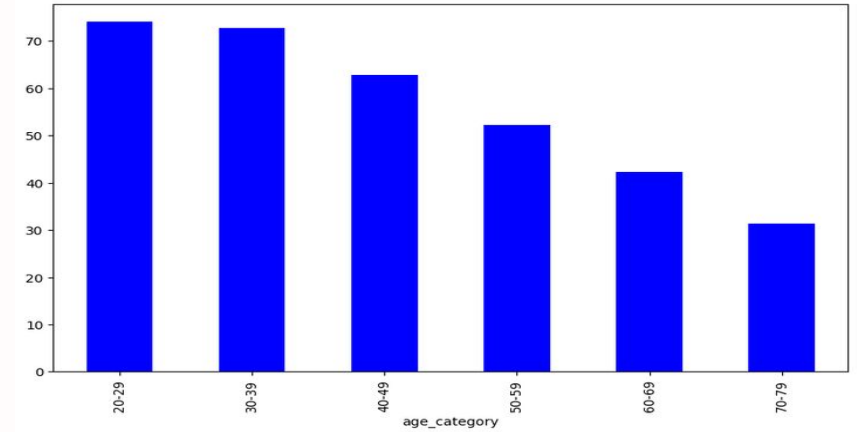


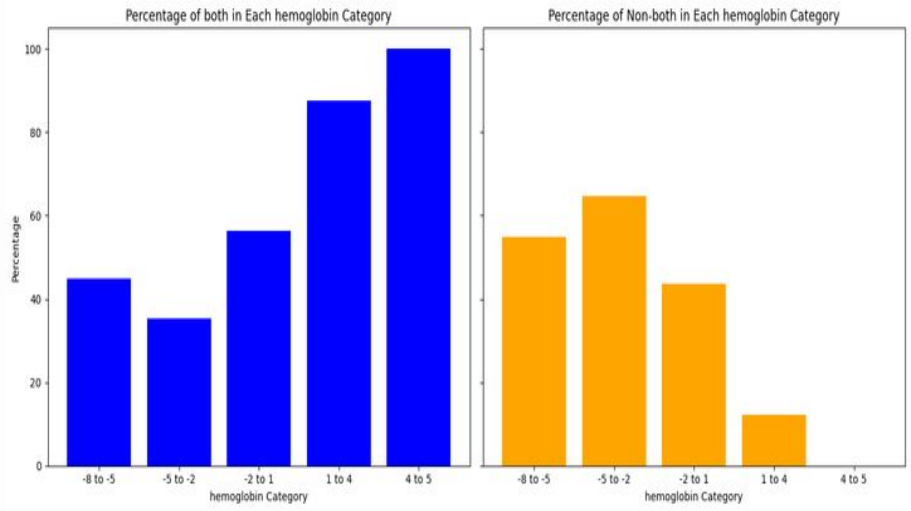
DRINKING

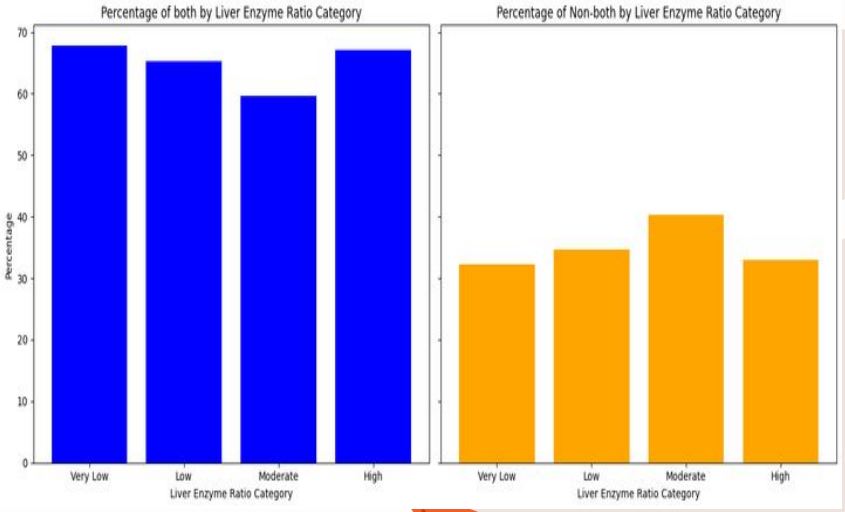

Percentage of Smokers by Liver Damage Score Category
Percentage
0
10
20
30
40
50
60
70
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category
Percentage of Non-Smokers by Liver Damage Score Category
Very Low
Low
Moderate
High
Very High
Extremely High
Liver Damage Score Category

Drinking


Percentage of people doing smoking or drinking or both
Male
Female
Gender


20-29
30-39
40-49
50-59
60-69
70-79
age_category


Percentage of both in Each hemoglobin Category
Percentage of Non-both in Each hemoglobin Category
Percentage
-8 to -5
-5 to -2
-2 to 1
1 to 4
4 to 5
hemoglobin Category


Percentage of both by Liver Enzyme Ratio Category
Percentage of Non-both by Liver Enzyme Ratio Category
Percentage
Very Low
Low
Moderate
High
Liver Enzyme Ratio Category

Insights

Smoking and drinking behaviors show a sharp increase in likelihood between ages 20 and 30, with a significant decline after 45, likely due to lifestyle changes and increased responsibilities.

Hemoglobin levels are notably higher in smokers and drinkers, while liver damage scores are significantly elevated in drinkers compared to non-drinkers.

There is a high conditional probability that individuals who smoke are also likely to engage in drinking behavior.

Smokers and drinkers tend to have higher triglyceride levels compared to non-smokers and non-drinkers.

Baseline Accuracy

Drinkers: 50.19%

We assume everyone is a drinker

Smokers: 60.77%

We assume everyone is not a smoker

Drinkers and/or Smokers: 60.91%

We assume everyone drinks and/or smokes

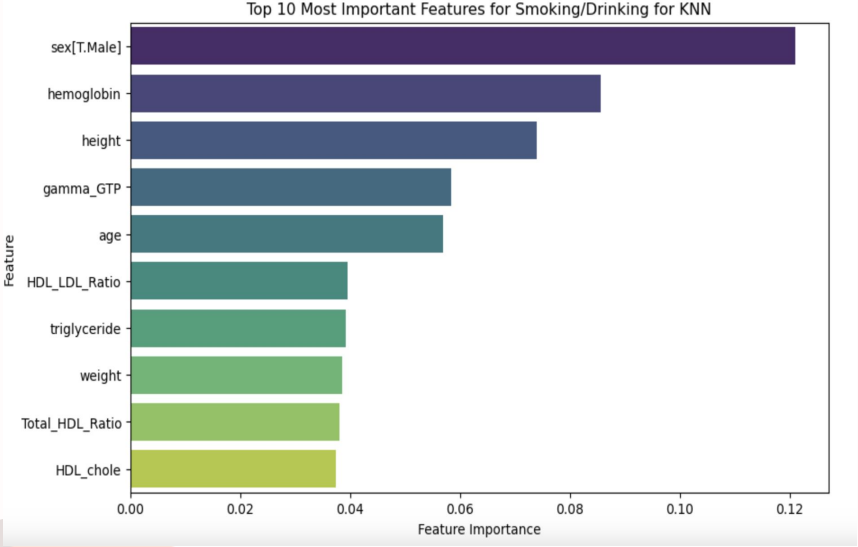
Model Results - Accuracy

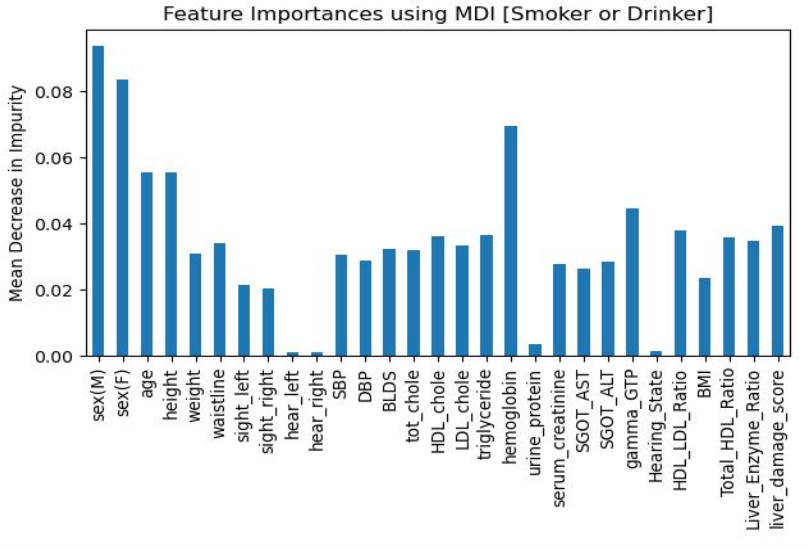
	No FE		FE		
Category	Drinker	Smoker	Drinker	Smoker	Drinker or Smoker
Model					
Logistic Regression	71.13%	78.26%	71.47%	79.04%	78.33%
KNN	68.44%	63.76%	66.42%	74.64%	73.40%
Naive Bayes	68.37%	64.68%	69.34%	72.98%	73.84%
Decision Trees	71.23%	80.96%	71.80%	80.96%	78.96%
Bagging	69.50%	78.40%	69.56%	78.37%	77.37%
Random Forest	72.59%	80.83%	72.49%	80.94%	79.24%

Variable Importance

Age, Hemoglobin, and Gamma-GTP consistently emerged as the most important across all models.

Liver Damage Score and Liver Enzyme Ratio were significant engineered features



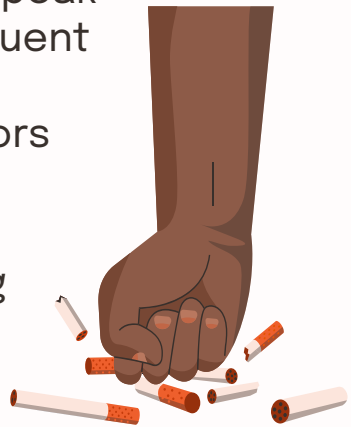


Conclusion

- Drinkers
  - Best Model: Random Forest
  - Improvement over Baseline: 22.40%
  - Accuracy: 72.49%
- Smokers
  - Best Model: Decision Tree
  - Improvement over Baseline: 20.19%
  - Accuracy: 80.94%
- Drinker and/or Smoker
  - Best Model: Random Forest
  - Improvement over Baseline: 18.33%
  - Accuracy: 79.24%

Insights

- Utilize age, hemoglobin levels, liver damage scores, Gamma-GTP, and liver enzyme ratios to accurately predict smoking and drinking behaviors.
- Recognize the high likelihood that smokers also engage in drinking, and integrate this correlation into underwriting models to enhance risk predictions and set more precise premiums.
- Adjust premiums and risk management strategies to account for peak smoking and drinking behaviors in younger adults and the subsequent decline observed in older individuals.
- Use elevated hemoglobin and liver damage scores as key indicators to refine risk assessments and adjust premiums accordingly.
- Shift from relying on self-reported data to leveraging predictive analytics for setting premiums, ensuring fair and accurate pricing based on actual risk profiles.



Thank you!