Identifying Key Predictors of Smoking Abstinence through Lasso Regression and Best Subset Modeling

Michael Stewart

Abstract

This project, in collaboration with Dr. George Papandonatos from Brown University's Biostatistics Department, explores factors impacting smoking cessation in adults with major depressive disorder (MDD). Individuals with MDD often smoke more heavily, face greater nicotine dependence, and experience more severe withdrawal, making cessation efforts particularly challenging. We evaluated the smoking cessation drug varenicline alongside behavioral therapies tailored to address these specific challenges. Data came from a randomized, placebo-controlled, 2x2 factorial trial comparing behavioral activation for smoking cessation (BASC) to standard treatment (ST), and varenicline to placebo, in a sample of 300 adults with current or past MDD. Earlier findings indicated that BASC did not significantly outperform ST.

In our analysis, we examined baseline characteristics as potential moderators of treatment effects on end-of-treatment abstinence, applying Lasso and Best Subset regression models to uncover key predictors. Results showed that participants on placebo were significantly less likely to quit smoking compared to those on varenicline. Higher tobacco dependence scores (FTCD) also predicted lower abstinence rates, highlighting the challenge of quitting for individuals with stronger dependencies. Non-Hispanic White participants were marginally more likely to achieve abstinence, and the Nicotine Metabolism Ratio (NMR) had a modest association with outcomes.

For model evaluation, we used AUC and calibration plots to assess both discrimination and calibration of the predictions. The Lasso model captured a range of associations, while the Best Subset model highlighted the most stable predictors, with treatment type and dependence level emerging as critical factors. Overall, these findings suggest that quitting interventions for MDD populations may benefit from an integrated approach combining pharmacologic and behavioral support.

1. Introduction

Quitting smoking presents unique challenges for individuals with major depressive disorder (MDD), who often face higher nicotine dependence and more intense withdrawal symptoms than the general population. These additional barriers make it crucial to understand which baseline factors might predict or influence treatment success. Demographic characteristics, depression severity, and tobacco dependence levels are some variables that could shape treatment outcomes, but their roles in smoking cessation for MDD populations are not fully clear.

To explore this, we used both Lasso and Best Subset regression models, taking advantage of their different strengths. Lasso regression, which applies variable selection and shrinkage, is well-suited to handling datasets with many predictors. By applying a penalty to reduce overfitting, Lasso identifies a comprehensive set of potential predictors without making the model overly complex.

Best Subset selection, on the other hand, focuses on finding the most stable predictors by evaluating combinations of variables and selecting the ones that best explain the outcome. Although computationally demanding, this approach provides a refined, interpretable model that can highlight the most meaningful predictors—a valuable aspect in clinical research, where pinpointing key influences can inform more targeted treatment strategies.

By combining Lasso's broad, exploratory approach with Best Subset's focused analysis, we aim to gain a thorough understanding of the factors influencing smoking cessation success in MDD populations. Together, these models help us identify key baseline characteristics and treatment effects that could shape more effective, tailored interventions for people with MDD looking to quit smoking.

2 Data Description and Preprocessing

2.1 Data Source and Study Design

The trial enrolled 300 adult smokers with current or past MDD, who were randomly assigned to one of four treatment groups: varenicline with behavioral activation for smoking cessation (BASC), varenicline with standard treatment (ST), placebo with BASC, or placebo with ST. BASC is a behavioral approach that aims to increase engagement in rewarding, non-smoking-related activities to counteract depressive symptoms, while ST focuses on general smoking cessation counseling without targeted behavioral activation. This design allowed for comparisons of the effects of varenicline versus placebo and BASC versus ST, both individually and in combination.

Table 1 presents a summary of the baseline variables for study participants, including demographic factors (age, sex, race/ethnicity), nicotine dependence indicators (such as the FTCD score and Nicotine Metabolism Ratio [NMR]), and psychological variables (including the Beck Depression Depression Inventory [BDI] score and readiness to quit).

Table 1: Participant Characteristics by Treatment Arm and Overall Sample, with Missing Data

	Treatment Groups				
Characteristic	**Overall**, N = 300	**BASC + Placebo**, N = 68	**BASC + Varenicline**, N = 83	**ST + Placebo**, N = 68	**ST + Varenicline**, N = 81
Age, Mean (SD)	50 (13)	51 (14)	50 (13)	50 (11)	49 (13)
Sex, n (%)					
Female	165 (55%)	38 (56%)	44 (53%)	39 (57%)	44 (54%)
Male	135 (45%)	30 (44%)	39 (47%)	29 (43%)	37 (46%)
Non-Hispanic White, n (%)	105 (35%)	24 (35%)	34 (41%)	22 (32%)	25 (31%)
Black, n (%)	157 (52%)	37 (54%)	37 (45%)	40 (59%)	43 (53%)
Hispanic, n (%)	18 (6.0%)	5 (7.4%)	4 (4.8%)	4 (5.9%)	5 (6.2%)
Income, n (%)					
$20,000–35,000	68 (23%)	16 (24%)	17 (20%)	14 (21%)	21 (26%)
$35,001–50,000	46 (15%)	8 (12%)	13 (16%)	14 (21%)	11 (14%)
$50,001–75,000	38 (13%)	12 (18%)	12 (14%)	8 (12%)	6 (7.4%)
Less than $20,000	110 (37%)	25 (37%)	30 (36%)	26 (38%)	29 (36%)
More than $75,000	35 (12%)	6 (8.8%)	10 (12%)	6 (8.8%)	13 (16%)
Unknown	3 (1.0%)	1 (1.5%)	1 (1.2%)	0 (0%)	1 (1.2%)
Education Level, n (%)					
College graduate	91 (30%)	19 (28%)	29 (35%)	17 (25%)	26 (32%)
Grade school	1 (0.3%)	1 (1.5%)	0 (0%)	0 (0%)	0 (0%)
High school graduate or GED	76 (25%)	23 (34%)	15 (18%)	11 (16%)	27 (33%)
Some college/technical school	116 (39%)	22 (32%)	32 (39%)	38 (56%)	24 (30%)
Some high school	16 (5.3%)	3 (4.4%)	7 (8.4%)	2 (2.9%)	4 (4.9%)
FTCD Score, Mean (SD)	5 (2)	5 (2)	5 (2)	5 (2)	5 (2)
Unknown	1	0	0	1	0
FTCD Score (5 mins), n (%)	138 (46%)	32 (47%)	33 (40%)	35 (51%)	38 (47%)
BDI Score, Mean (SD)	19 (11)	19 (12)	18 (11)	18 (11)	20 (12)
Cigarettes per day, Mean (SD)	15 (8)	16 (9)	16 (9)	15 (7)	14 (7)
Craving Total, Mean (SD)	7 (4)	7 (4)	7 (4)	7 (4)	7 (3)
Unknown	18	1	3	8	6
Hedonic Sum (Negative), Mean (SD)	23 (20)	23 (20)	23 (19)	21 (20)	23 (19)
Hedonic Sum (Positive), Mean (SD)	25 (19)	28 (22)	22 (17)	27 (20)	25 (19)
Shaps Score, Mean (SD)	2 (3)	2 (3)	2 (3)	3 (3)	2 (3)
Unknown	3	2	0	1	0
Other Diagnoses, n (%)	133 (44%)	35 (51%)	30 (36%)	28 (41%)	40 (49%)
Antidepressant Medication, n (%)	82 (27%)	28 (41%)	24 (29%)	15 (22%)	15 (19%)
Current Major Depression Episode, n (%)	147 (49%)	32 (47%)	40 (48%)	31 (46%)	44 (54%)
Nicotine Metabolism Ratio, Mean (SD)	0.36 (0.23)	0.34 (0.18)	0.38 (0.25)	0.37 (0.27)	0.36 (0.21)
Unknown	21	7	3	2	9
Only Menthol, n (%)	178 (60%)	40 (59%)	48 (59%)	43 (64%)	47 (58%)
Unknown	2	0	1	1	0
Readiness to Quit, Mean (SD)	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Unknown	17	4	5	4	4

[1] Mean (SD) for continuous variables; n (%) for categorical variables.

2.2 Handling Missing Data

To address missing data, we first examined the extent and patterns of missingness across baseline variables. As summarized in **Table 1**, certain variables exhibited higher rates of missing data, including income, FTCD score (a measure of nicotine dependence), craving total, anhedonia score, Nicotine Metabolism Ratio (NMR), exclusive menthol use, and readiness to quit.

To address missing data, we used multiple imputation, generating five imputed datasets. We chose multiple imputation over complete case analysis to address missing data. A complete case analysis would have removed approximately 20% of the sample, reducing statistical power and potentially introducing bias if the missingness was not completely random (MCAR).

Additionally, due to the unique value of one participant who only completed grade school, this case was removed

from the analysis as it was causing instability in the Best Subset model.

After analyzing each imputed dataset separately, we applied Rubin's Rules to pool the results to get reliable, stable estimates of coefficients and standard errors.

2.3 Encoding Variables and Transformations

The two primary treatment variables, Varenicline (Var) and Behavioral Activation (BA), were coded as binary dummy variables, indicating whether a participant received varenicline (1 for varenicline, 0 for placebo) or behavioral activation (1 for BASC, 0 for standard treatment). We created an additional single treatment arm variable to represent the four unique combinations of treatment conditions: BASC + Placebo, BASC + Varenicline, ST + Placebo, and ST + Varenicline. In our regression analysis, we used this variable. This combined variable enabled efficient comparison across all treatment groups.

Additionally, we evaluated the distributions of continuous variables and considered log-transformations to reduce skewness. However, transformations did not significantly improve the distributions, and given that Lasso and Best Subset regression methods are generally robust to skewness, we decided not to apply log transformations. This helps with keeping models straightforward and easy to interpret.

3. Exploratory Analysis

The exploratory data analysis phase involved several analyses to understand variable relationships and assess their significance as predictors of smoking cessation.

The correlation heatmap (**Figure 1**) offered a visual representation of relationships between continuous variables. We wanted to explore if any variables had some collinearity. This is important because multi-collinearity can impact the stability of variable selection in Lasso and Best Subset modeling. The highest Pearson correlation was observed between FTCD score and cigarettes per day (cpd_ps), with a correlation of 0.51, indicating a moderate positive association. Another notable correlation was between bdi_score_w00 (a measure of baseline depression) and shaps_score_pq1 (anhedonia score), with a correlation of 0.41, suggesting a moderate association between depression and anhedonia. Overall, multicollinearity was low enough to support the inclusion of all predictors in the Lasso and Best Subset models without significant issues.

Figure 1: Correlation Heatmap

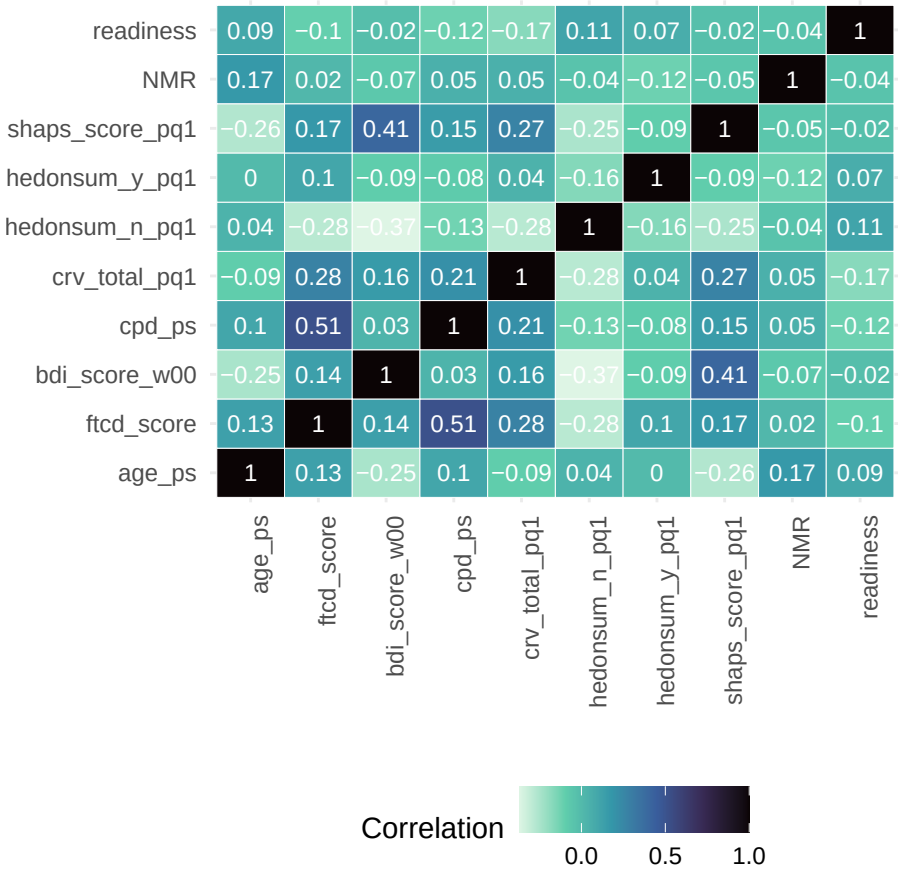
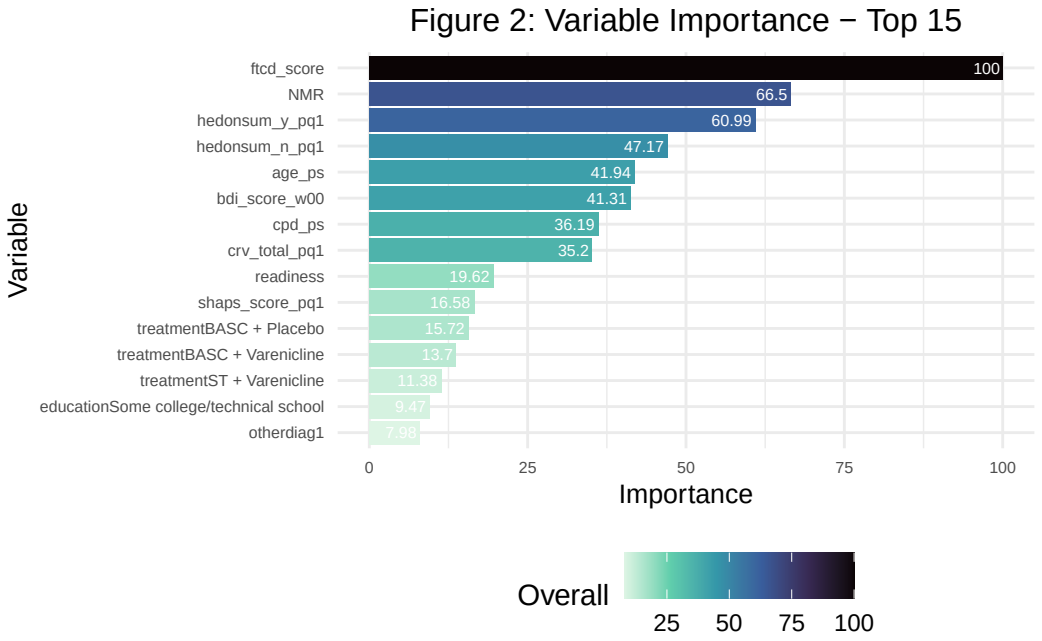


Table 2: Significant Chi-Square Test Results

Variable 1	Variable 2	X-squared	df	p-value
Black	NHW	173.10	1	< 2e-16
edu	inc	88.52	12	9.54e-14
Black	Only.Menthol	68.90	1	< 2e-16
NHW	Only.Menthol	52.89	1	3.53e-13
Black	edu	38.09	3	2.70e-08
inc	Only.Menthol	35.95	4	2.96e-07
Black	inc	34.91	4	4.86e-07
edu	Only.Menthol	27.21	3	5.33e-06
mde_curr	otherdiag	25.33	1	4.83e-07
inc	NHW	23.55	4	9.81e-05
abst	Var	23.32	1	1.37e-06
edu	NHW	20.88	3	1.11e-04
inc	mde_curr	14.83	4	5.07e-03
Black	Hisp	11.46	1	7.13e-04
edu	ftcd.5.mins	9.76	3	2.07e-02
Hisp	NHW	8.65	1	3.27e-03
Hisp	Only.Menthol	6.89	1	8.69e-03
NHW	sex_ps	6.62	1	1.01e-02
antidepmed	BA	6.59	1	1.02e-02
abst	NHW	5.10	1	2.39e-02
Black	sex_ps	4.77	1	2.89e-02
antidepmed	NHW	4.01	1	4.53e-02

We conducted a chi-square test (**Table 2**) to examine associations between categorical variables, including treatment arms, race/ethnicity, education levels, and income brackets. Significant associations emerged, such as between education and income levels ($\chi^2 = 88.52$, $p < 0.001$) and between race/ethnicity and menthol cigarette use ($\chi^2 = 52.89$, $p < 0.001$). These findings suggest that demographic and socioeconomic factors may influence smoking cessation outcomes, potentially affecting responses to treatment.

Although the Lasso and Best Subset models independently selected variables based on predictive power, these associations provided valuable context for interpreting results. For example, if demographic or socioeconomic factors, such as race or income, appeared as important predictors in our models, knowing that these factors are associated with other variables (like menthol cigarette use) helped us understand why they might influence smoking cessation outcomes. This context allows us to interpret the results with an understanding of the broader social or behavioral patterns that may impact abstinence success.



4. Methods

For our main analyses, we applied two regression modeling approaches: Lasso and Best Subset. These models were chosen to identify key predictors of abstinence and provide robust variable selection through different methods.

4.1 Lasso Regression

To identify key predictors of smoking abstinence, we applied Lasso regression because of its ability to perform variable selection by shrinking the coefficients of less important predictors toward zero. We began by splitting each imputed dataset into an 80/20 train-test split. We fit a Lasso model for each imputed dataset, performing 5-fold cross-validation to select the optimal penalty parameter, .We hose the smallest value as our optimal parameter. This process was repeated across all five imputed datasets, resulting in five sets of coefficient estimates. To obtain pooled coefficient estimates, we applied Rubin's Rules, calculating the mean coefficient for each predictor across the imputed datasets and combining within and between variances to get the pooled standard errors.

4.2 Best Subset Selection

To complement our Lasso analysis, we applied Best Subset regression, which sorts through all possible subsets of predictors to identify the combination that provides the best predictive performance. As with Lasso, we used an 80/20 train-test split for each imputed dataset. We ran Best Subset regression on each of the five imputed datasets. The model was fit using the L0Learn package with logistic loss function. We applied 5-fold cross-validation within each imputed dataset to find the optimal combination of tuning the gamma and lambda parameters that minimized cross-validation error. After selecting the optimal parameters for each imputed dataset, we extracted the coefficients for the chosen subset of predictors, obtaining five sets of coefficient estimates, one for each imputed dataset.Similar to Lasso, we combined results using Rubin's Rules.

5. Results

5.1 Lasso Results

Table 3 shows the non-zero coefficients selected by Lasso, showing the variables with significant associations in the model. Notably, NHW (Non-Hispanic White) has the highest positive coefficient (10.52, SE = 0.23), suggesting a significant impact on the outcome. NMR also shows a positive coefficient (0.27, SE = 0.66), but its relatively high standard error makes this effect less certain. Among treatment variables, ST + Varenicline has a negligible positive coefficient (0.001, SE = 0.04), indicating little effect. In contrast, treatment ST + Placebo and treatment BASC + Placebo have large negative coefficients (-0.73 and -1.11, with SEs of 0.07 and 0.06, respectively), suggesting these treatment combinations are strongly associated with lower probabilities of abstinence. Treatment ST + Placebo and treatment BASC + Placebo show large negative associations, while ST + Varenicline has a near-zero effect, suggesting different treatment combinations may have varying impacts on the outcome.

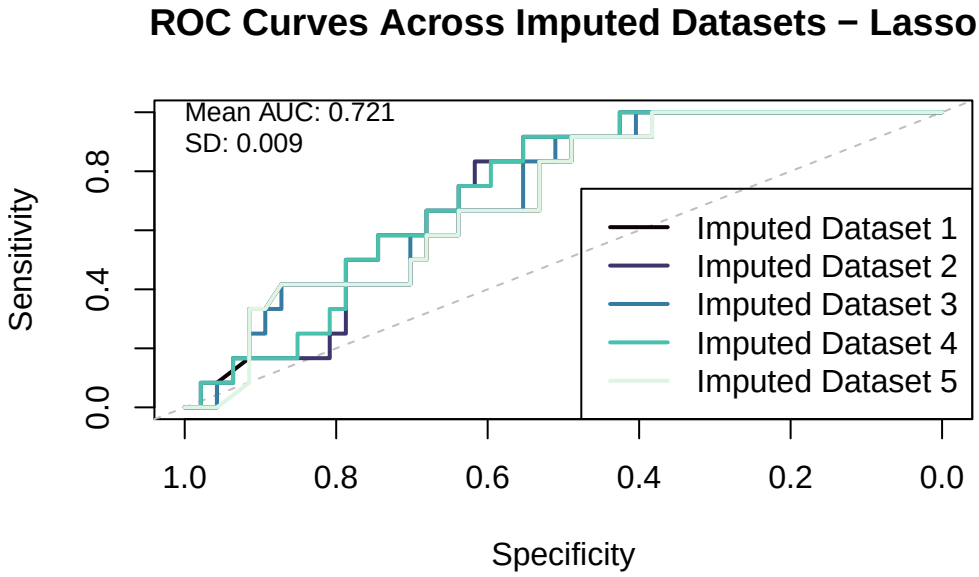
Moreover, the negative associations for some college or technical school education (educationSome college/technical school), current major depressive episode (mde_curr1), other psychiatric diagnoses (otherdiag1), and FTCD score (a measure of nicotine dependence) suggest that these factors may reduce the likelihood of abstinence, which seems theoretically sound.

Table 3: Lasso Model Selected Variables (Non-Zero Coefficients)

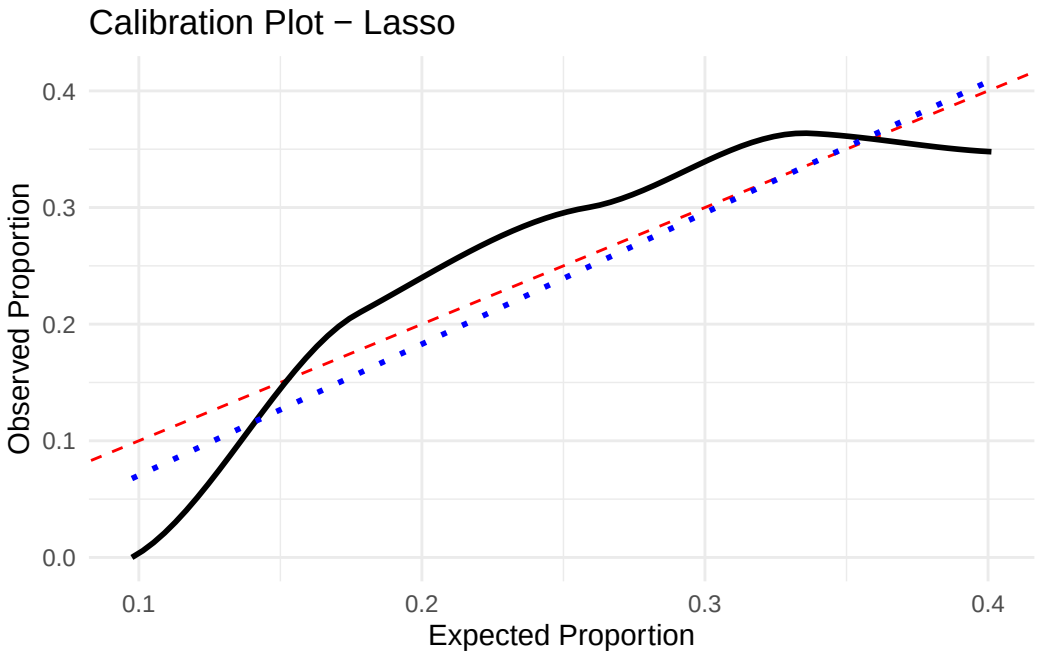
	Variable	Mean	SE
13	NHW1	0.5205052	0.2270132
30	NMR	0.2696027	0.6551439
10	incomeUnknown	0.0741871	0.2203237
5	treatmentST + Varenicline	0.0010038	0.0448740
17	educationSome college/technical school	-0.1003259	0.1210112
29	mde_curr1	-0.1086615	0.0796562
27	otherdiag1	-0.1126829	0.0879005
19	ftcd_score	-0.1443972	0.0425614
4	treatmentST + Placebo	-0.7255700	0.0708039
2	treatmentBASC + Placebo	-1.1083130	0.0636238

The ROC curves for the Lasso model across the five imputed datasets indicate a consistent performance, with each curve closely following a similar trajectory. The mean AUC of 0.721 and a standard deviation of 0.009 suggest moderate predictive accuracy and low variability between imputations.

These results indicate that the Lasso model has stable sensitivity and specificity across the different datasets, with its predictive ability being relatively consistent regardless of data imputation. The small standard deviation also reinforces the reliability of the AUC estimate across multiple imputations, providing confidence in the model's predictive power for the chosen variables.



In this calibration plot for the Lasso model below, using 5 bins, the black loess curve shows underestimation at both low and high expected proportions, while there is overestimation in the middle range The blue dotted line represents a linear fit, capturing the overall trend of observed vs. expected proportions with a slightly different slope from the ideal line. The linear fit closely aligns with the overall trend. However there are some deviations in calibration accuracy at the extremes.



5.2 Best Subset Results

The Best Subset model provides some insights into predictors of smoking abstinence. In **Table 4**, The "BASC + Placebo" treatment group shows a strong negative impact on abstinence, with a large negative coefficient (-2.06) and a low standard error (0.10), indicating a stable, significant association. Similarly, the "ST + Placebo" group also has a negative association with abstinence (-1.42), though slightly smaller in effect size and with a bit more variability (standard error of 0.15).

The FTCD score, representing nicotine dependence, shows a negative association with abstinence (-0.24) and has a low standard error (0.05), which makes theoretical sense because higher dependence most likely lowers abstinence probability.

Being non-Hispanic White (NHW1) is positively linked to abstinence (0.92), suggesting some demographic influence,

though with a higher standard error (0.29), indicating more variability in this effect.

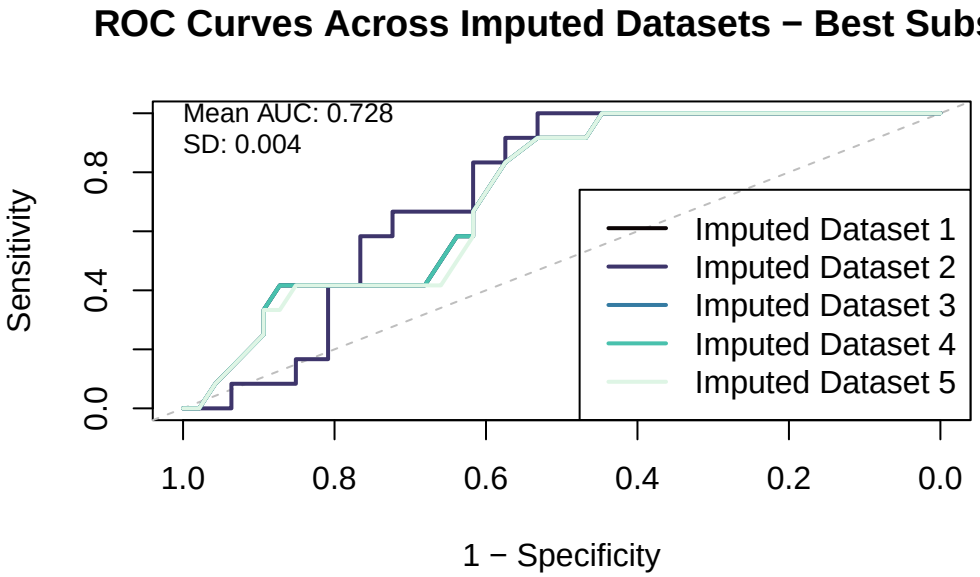
The nicotine metabolism ratio (NMR) has a positive coefficient (0.26) but a high standard error (0.95), pointing to potential variability / inconsistencies across imputations.

Overall, while the findings generally align with expected effects (like the impact of nicotine dependence and treatment with placebo), the high standard errors for some predictors, such as NMR, show some instbaility in prediction.

Table 4: Best Subset Model Selected Variables (Non-Zero Coefficients)

Variable	Mean	SE
NHW1	0.9153616	0.2917645
NMR	0.2554714	0.9500191
ftcd_score	-0.2375431	0.0516465
treatmentST + Placebo	-1.4197190	0.1516206
treatmentBASC + Placebo	-2.0577662	0.0989415

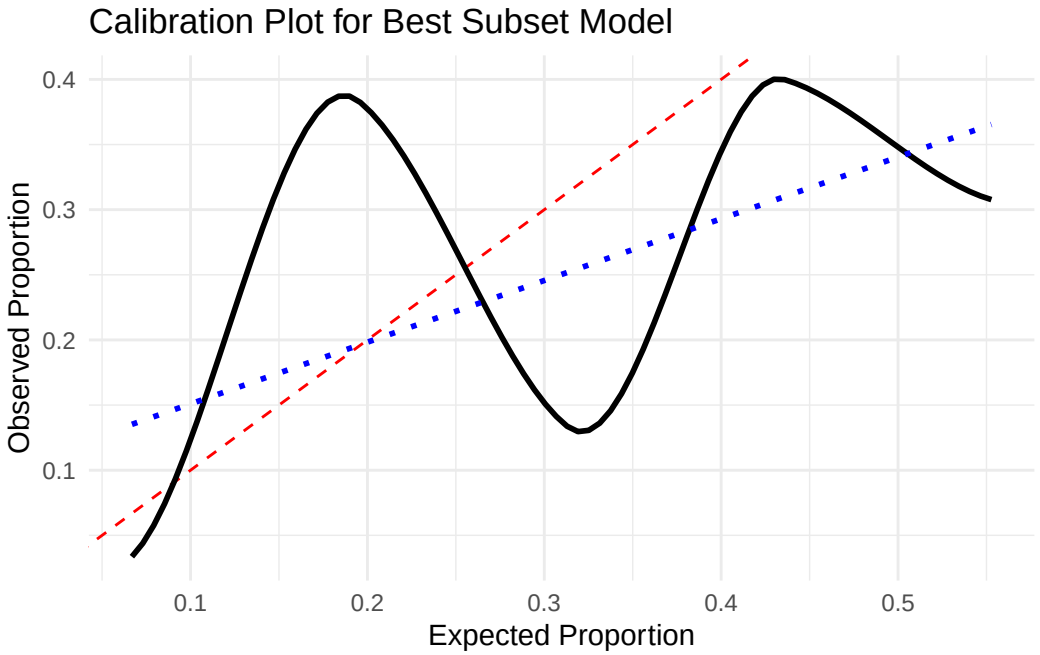
The ROC curves for the Best Subset model across the five imputed datasets indicate consistent performance, with an average AUC of 0.728 and a low standard deviation (0.004). This suggests that the model's predictive accuracy and discrimination is stable and reliable across multiple imputations.



The calibration plot for the Best Subset model displays notable deviations from the ideal line, with significant deviations in the observed proportions (black curve). The plot reveals a pattern of underestimation at lower expected proportions, overestimation in the mid-range, and another round of underestimation at higher values.

These oscillations suggest that the Best Subset model may be potentially overfitting specific data points. The blue dotted line, representing a linear fit, diverges from the ideal but shows a more stable slope, suggesting that there is a moderate relationship between predicted and observed values, but not a strong one.

This plot highlights areas for improvement in the model's calibration. It suggests that further model tuning may be necessary to achieve better alignment between predicted probabilities and observed outcomes.



5.3 Comparison

Both models identified Non-Hispanic White (NHW) status, NMR (nicotine metabolism ratio), FTCD score, and the treatment conditions ST + Placebo and BASC + Placebo as significant variables, suggesting these factors play a consistent role in predicting abstinence. Lasso selected additional variables that the Best Subset model did not, such as income (Unknown), current Major Depressive Episode (MDE), education level (Some college/technical school), and other diagnoses (Otherdiag).

The ROC curves and AUC values for both models demonstrated similar predictive strength, with Lasso achieving a mean AUC of 0.721 (SD: 0.009) and Best Subset slightly higher at 0.728 (SD: 0.004). Calibration plots for each model indicated a pattern of underestimation at the extremes of expected probabilities and a tendency to overestimate in the mid-range. Lasso exhibited tighter calibration overall, perhaps due to its inclusion of additional variables.

6. Conclusion and Limitations

Our analysis identified treatment type and nicotine dependence (ftcd_score) as key predictors of smoking cessation in adults with MDD. Both Lasso and Best Subset models consistently highlighted these variables, aligning with the correlation heatmap, which showed their association with cessation outcomes. However, chi-square tests emphasized demographic factors like race, education, and menthol use rather than treatment type, suggesting these characteristics are related to smoking behavior but may have limited predictive power for cessation success.

This project has some limitations. Lasso's regularization can exclude weaker predictors, potentially missing subtle effects, while Best Subset risks overfitting by testing all combinations. The calibration plots, particularly for the Best Subset model, also revealed significant deviations that suggest overfitting. With a relatively high number of predictors and a limited sample size, the risk of overfitting was heightened, as small datasets often struggle with generalizability. These limitations suggest that future work might benefit from external validation sets or incorporating non-linear models to improve stability and generalizability.

References

Hitsman, B., Papandonatos, G. D., Gollan, J. K., Huffman, M. D., Niaura, R., Mohr, D. C., Veluz-Wilkins, A. K., Lubitz, S. F., Hole, A., Leone, F. T., Khan, S. S., Fox, E. N., Bauer, A. M., Wileyto, E. P., Bastian, J., & Schnoll, R. A. (2023). Efficacy and safety of combination behavioral activation for smoking cessation and varenicline for treating tobacco dependence among individuals with current or past major depressive disorder: A 2 x 2 factorial, randomized, placebo-controlled trial. Addiction (Abingdon, England), 118(9), 1710–1725. https://doi.org/10.1111/add.16209

Code Appendix

```
knitr::opts_chunk$set(warning = FALSE,
                      message = FALSE,
                      echo = FALSE,
                      fig.align = "center")

library(tidyverse)
library(knitr)
library(kableExtra)
library(mice)
library(corrplot)
library(gtsummary)
library(RColorBrewer)
library(reshape2)
library(gridExtra)
library(glmnet)
library(randomForest)
library(caret)
library(LOLearn)
library(pROC)

project2 <- read.csv("~/Downloads/project2.csv")


# define treatment categories
project2 <- project2 %>%
  mutate(
    treatment = case_when(
      Var == 1 & BA == 1 ~ "BASC + Varenicline",
      Var == 0 & BA == 1 ~ "BASC + Placebo",
      Var == 1 & BA == 0 ~ "ST + Varenicline",
      Var == 0 & BA == 0 ~ "ST + Placebo"
    )
  )

# recode income, education, and sex levels
project2 <- project2 %>%
  mutate(
    income = case_when(
      inc == 1 ~ "Less than $20,000",
      inc == 2 ~ "$20,000-35,000",
      inc == 3 ~ "$35,001-50,000",
      inc == 4 ~ "$50,001-75,000",
      inc == 5 ~ "More than $75,000",
      TRUE ~ "Unknown"
    ),
    education = case_when(
      edu == 1 ~ "Grade school",
      edu == 2 ~ "Some high school",
      edu == 3 ~ "High school graduate or GED",
      edu == 4 ~ "Some college/technical school",
      edu == 5 ~ "College graduate",
      TRUE ~ "Unknown"
    ),
    sex = case_when(
      sex_ps == 1 ~ "Male",
```

```
      sex_ps == 2 ~ "Female",
      TRUE ~ "Unknown")
  )

# summary table with baseline variables
baseline_table <- project2 %>%
  select(
    treatment, age_ps, sex, NHW, Black, Hisp, income, education,
    ftcd_score, ftcd.5.mins, bdi_score_w00, cpd_ps, crv_total_pq1,
    hedonsum_n_pq1, hedonsum_y_pq1, shaps_score_pq1, otherdiag,
    antidepmed, mde_curr, NMR, Only.Menthol, readiness
  ) %>%
  tbl_summary(
    by = treatment,
    label = list(
      age_ps ~ "Age",
      sex ~ "Sex",
      NHW ~ "Non-Hispanic White",
      Black ~ "Black",
      Hisp ~ "Hispanic",
      income ~ "Income",
      education ~ "Education Level",
      ftcd_score ~ "FTCD Score",
      ftcd.5.mins ~ "FTCD Score (5 mins)",
      bdi_score_w00 ~ "BDI Score",
      cpd_ps ~ "Cigarettes per day",
      crv_total_pq1 ~ "Craving Total",
      hedonsum_n_pq1 ~ "Hedonic Sum (Negative)",
      hedonsum_y_pq1 ~ "Hedonic Sum (Positive)",
      shaps_score_pq1 ~ "Shaps Score",
      otherdiag ~ "Other Diagnoses",
      antidepmed ~ "Antidepressant Medication",
      mde_curr ~ "Current Major Depression Episode",
      NMR ~ "Nicotine Metabolism Ratio",
      Only.Menthol ~ "Only Menthol",
      readiness ~ "Readiness to Quit"
    ),
    statistic = c(all_continuous() ~ "{mean} ({sd})",
                  all_categorical() ~ "{n} ({p}%)"),
    type = list(readiness ~ "continuous"),
    missing = "ifany"
  ) %>%
  add_overall() %>%
  add_stat_label() %>%
  modify_spanning_header(
    all_stat_cols() ~ "**Treatment Groups**"
  ) %>%
  modify_footnote(
    all_stat_cols() ~ "Mean (SD) for continuous variables; n (%) for categorical variables."
  ) %>%
  as_kable_extra(
    booktabs = TRUE,
    caption = "Participant Characteristics by Treatment Arm and Overall Sample, with Missing Data"
  ) %>%
  kable_styling(
    latex_options = c("hold_position", "scale_down", "striped"),
```

```
    full_width = FALSE,
    position = "center"
  )

baseline_table


project2 <- project2[project2$education != "Grade school", ]

# continuous and categorical variables
continuous_vars <- c("age_ps", "ftcd_score", "bdi_score_w00", "cpd_ps",
                     "crv_total_pq1", "hedonsum_n_pq1", "hedonsum_y_pq1",
                     "shaps_score_pq1", "NMR", "readiness")
categorical_vars <- c("abst", "Var", "BA", "inc", "edu", "sex_ps", "NHW", "Black",
                      "Hisp", "ftcd.5.mins", "otherdiag",
                      "antidepmed", "mde_curr", "Only.Menthol")

# make categorical variables factors
project2 <- project2 %>%
  mutate(across(all_of(categorical_vars), as.factor))

# continuous variables and make it a matrix
continuous_vars <- project2 %>%
  select(-id) %>%
  select_if(is.numeric) %>%
  as.matrix()  # Convert to matrix

# correlation matrix
cor_matrix <- cor(continuous_vars, use = "complete.obs")
cor_melted <- melt(cor_matrix)

# correlation heatmap
ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(value, 2)), color = "white", size = 3) +  # Add correlation values as text
  scale_fill_viridis_c(option = "mako", direction = -1, name = "Correlation") +
  labs(title = "Figure 1: Correlation Heatmap", x = "", y = "") +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    axis.text.x = element_text(angle = 90, hjust = 1),
    plot.title = element_text(hjust = 0.5)
  )



# chi-square test funciton
chi_square <- function(var1, var2, data) {
  contingency_table <- table(data[[var1]], data[[var2]])
  test_result <- chisq.test(contingency_table)
  data.frame(
    Variable1 = var1,
    Variable2 = var2,
    X_squared = test_result$statistic,
    df = test_result$parameter,
    p_value = test_result$p.value
```

```
  )
}

chi_square_results <- expand.grid(var1 = categorical_vars,
                                  var2 = categorical_vars,
                                  stringsAsFactors = FALSE) %>%
  filter(var1 < var2) %>%  # Exclude self-pairs
  rowwise() %>%
  do(chi_square(.$var1, .$var2, data = project2)) %>%  # run for each pair
  bind_rows() %>%
  ungroup()


signif_results <- chi_square_results %>%
  filter(p_value < 0.05) %>%         # keep only signif results
  rowwise() %>%
  mutate(
    pair_key = paste(sort(c(Variable1, Variable2)), collapse = "_")
  ) %>%
  distinct(pair_key, .keep_all = TRUE) %>% # no self pairs or repeats
  select(-pair_key) %>%
  ungroup() %>%
  arrange(desc(X_squared)) %>%
  mutate(
    X_squared = round(X_squared, 2),
    p_value = format.pval(p_value, digits = 3, scientific = TRUE)
  )

kable(signif_results, format = "latex", booktabs = TRUE, align = "lcccc",
      col.names = c("Variable 1", "Variable 2", "X-squared", "df", "p-value"),
      caption = "Significant Chi-Square Test Results") %>%
  kable_styling(latex_options = c("hold_position", "striped"))


project2_clean <- na.omit(project2)
set.seed(1)

# cv
cv_control <- trainControl(method = "cv", number = 5, verboseIter = F)
formula <- abst ~ treatment + income + sex + age_ps + NHW + Black + Hisp +
                    education + ftcd_score + ftcd.5.mins + bdi_score_w00 +
                    cpd_ps + crv_total_pq1 + hedonsum_n_pq1 + hedonsum_y_pq1 +
                    shaps_score_pq1 + otherdiag + antidepmed + mde_curr +
                    NMR + Only.Menthol + readiness - 1


# 5-fold cross-validation on random forest
rf_cv <- train(
  formula,
  data = project2_clean,
  method = "rf",
  trControl = cv_control,
  ntree = 500
)

#  convert var imp to a data frame
```

```r
varimp_data <- varImp(rf_cv)$importance
varimp_data$Variable <- rownames(varimp_data)  # Add variable names as a column

top_varimp_data <- varimp_data %>%
  arrange(desc(Overall)) %>%
  head(15)

ggplot(top_varimp_data, aes(x = Overall, y = reorder(Variable, Overall))) +
  geom_bar(stat = "identity", aes(fill = Overall)) +
  geom_text(aes(label = round(Overall, 2)),
            hjust = 1.1, vjust = 0.5, size = 2, color = "white") +
  scale_fill_viridis_c(option = "mako", direction = -1) +
  labs(
    title = "Figure 2: Variable Importance - Top 15",
    x = "Importance",
    y = "Variable"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, size = 12),
    axis.text = element_text(size = 6),
    axis.title = element_text(size = 10)
  )


imputed_data <- mice(project2, m = 5, maxit = 50, seed = 1, print = F)

set.seed(1)
project2$treatment <- as.factor(project2$treatment)


m <- 5
lasso_coef_estimates <- list()
lasso_optimal_lambdas <- list()
lasso_test_predictions <- list()

#  train-test split
trainIndex <- createDataPartition(imputed_data[[1]]$abst, p = 0.8, list = FALSE)

for (i in 1:m) {
  completed_data <- complete(imputed_data, i)

  # tes,train
  train_data <- completed_data[trainIndex, ]
  test_data <- completed_data[-trainIndex, ]

  # model matrix
  X_train <- model.matrix(formula, data = train_data)
  Y_train <- train_data$abst
  X_test <- model.matrix(formula, data = test_data)
  Y_test <- test_data$abst

  # cv lasso model
  cv_fit <- cv.glmnet(X_train, Y_train, alpha = 1, family = "binomial")
```

```r
  # store coeffs
  lasso_coef_estimates[[i]] <- coef(cv_fit, s = "lambda.min")

  # test set predictions (for roc)
  lasso_test_predictions[[i]] <- predict(cv_fit, newx = X_test, s = "lambda.min", type = "response")
}

# data frame to store pooled results
lasso_pooled_results <- data.frame(Variable = rownames(lasso_coef_estimates[[1]]),
                                   Mean = NA, SE = NA)

# calcualte mean and standard error for each coefficient
for (var in lasso_pooled_results$Variable) {
  coefs <- sapply(lasso_coef_estimates, function(x) as.numeric(x[var, 1]))

  lasso_pooled_results[lasso_pooled_results$Variable == var, "Mean"] <- mean(coefs, na.rm = TRUE)

  # Rubin's Rules for standard error calculation
  se_within <- sqrt(mean((coefs - mean(coefs))^2))
  se_between <- var(coefs, na.rm = TRUE)
  pooled_se <- sqrt(se_within + (1 + 1/m) * se_between)

  lasso_pooled_results[lasso_pooled_results$Variable == var, "SE"] <- pooled_se
}

# get non zero and order variables
lasso_selected_vars <- lasso_pooled_results[lasso_pooled_results$Mean != 0 &
                                              lasso_pooled_results$Variable != "(Intercept)", ]
lasso_selected_sorted <- lasso_selected_vars[order(-lasso_selected_vars$Mean), ]

# lasso table
lasso_table <- kable(lasso_selected_sorted, caption = "Lasso Model Selected Variables (Non-Zero Coefficient
  kable_styling(full_width = F, font_size = 12)

lasso_table


library(viridis)

# store auc
auc_values <- numeric(m)

# plot
plot(0, 0, type = "n", xlim = c(1, 0), ylim = c(0, 1),
     xlab = "Specificity", ylab = "Sensitivity",
     main = "ROC Curves Across Imputed Datasets - Lasso")
abline(a = 1, b = -1, col = "grey", lty = 2)

# loop to calculate and plot ROC curves
for (i in 1:m) {
  # Calculate ROC curve
  roc_curve <- roc(Y_test, lasso_test_predictions[[i]])
  auc_values[i] <- auc(roc_curve)  # Store AUC for this dataset
  plot(roc_curve, col = viridis(m, option = "mako")[i], lwd = 2, add = TRUE)
}
```

```r
# mean and standard deviation of AUC
mean_auc <- mean(auc_values)
sd_auc <- sd(auc_values)

text(x = 1, y = 0.95, labels = paste("Mean AUC:", round(mean_auc, 3), "\nSD:",
                                     round(sd_auc, 3)),
     adj = 0, col = "black", cex = 0.8)
legend("bottomright", legend = paste("Imputed Dataset", 1:m),
       col = viridis(m, option = "mako"), lwd = 2)



# prepare df for calibration data
calibration_data <- do.call(rbind, lapply(1:m, function(i) {
  data.frame(
    predicted = as.vector(lasso_test_predictions[[i]]),
    observed = as.numeric(Y_test) - 1  # Convert factor to numeric for binary outcomes
  )
}))

# bin the predictions
calibration_data <- calibration_data %>%
  mutate(bin = cut(predicted, breaks = 5)) %>%
  group_by(bin) %>%
  summarize(
    expected = mean(predicted),
    observed = mean(observed),
    se = sqrt((observed * (1 - observed)) / n())
  )

# create plot
ggplot(calibration_data, aes(x = expected, y = observed)) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  geom_smooth(method = "loess", color = "black", se = TRUE, fill = "gray") +
  geom_smooth(method = "lm", color = "blue", linetype = "dotted", se = F) +
  labs(
    x = "Expected Proportion",
    y = "Observed Proportion",
    title = "Calibration Plot - Lasso"
  ) +
  theme_minimal() +
    theme(legend.position = "bottom")


# set up train-test split (80% train, 20% test)
set.seed(1)  # for reproducibility
trainIndex <- createDataPartition(imputed_data[[1]]$abst, p = 0.8, list = FALSE)

# initialize lists for predictions, AUC values, and coefficient estimates
best_subset_test_predictions <- list()
best_subset_coef_estimates <- list()  # store coefficients for each imputed dataset
Y_test <- NULL  # store observed test outcomes once

# loop through each imputed dataset
for (i in 1:m) {
  completed_data <- complete(imputed_data, i)
```

```
  # split data into train and test sets
  train_data <- completed_data[trainIndex, ]
  test_data <- completed_data[-trainIndex, ]

  # prepare model matrix and outcome for training and testing
  X_train <- model.matrix(formula, data = train_data)
  Y_train <- train_data$abst
  X_test <- model.matrix(formula, data = test_data)
  Y_test <- test_data$abst

  # fit best subset model with cross-validation
  best_sub_model <- L0Learn.cvfit(X_train, Y_train, loss = "Logistic",
                                  nFolds = 5, maxIter = 100)

  # get optimal gamma and lambda
  min_ind <- which(best_sub_model$cvMeans[[1]] == min(best_sub_model$cvMeans[[1]]),
                   arr.ind = TRUE)
  gamma_min <- best_sub_model$fit$gamma[[min_ind[2]]]
  lambda_min <- best_sub_model$fit$lambda[[min_ind[2]]][min_ind[1]]

  # extract coefficients at optimal parameters
  coef_output <- as(coef(best_sub_model, gamma = gamma_min, lambda = lambda_min),
                    "matrix")
  rownames(coef_output) <- c("(Intercept)", colnames(X_train))
  best_subset_coef_estimates[[i]] <- coef_output  # store coefficients

  # make predictions on test set
  predictions <- predict(best_sub_model, newx = X_test, gamma = gamma_min,
                         lambda = lambda_min, type = "response")
  best_subset_test_predictions[[i]] <- predictions
}

# combine coefficients across imputations
best_subset_coef_matrix <- do.call(cbind, best_subset_coef_estimates)

# pool results using rubin's rules
variable_names <- rownames(best_subset_coef_estimates[[1]])
best_subset_pooled_results <- data.frame(Variable = variable_names,
                                         Mean = NA, SE = NA)

for (var in variable_names) {
  coefs <- best_subset_coef_matrix[var, ]
  best_subset_pooled_results[best_subset_pooled_results$Variable == var, "Mean"] <- mean(coefs, na.rm = TRU

  # calculate standard error using rubins rules
  se_within <- sqrt(mean((coefs - mean(coefs, na.rm = TRUE))^2, na.rm = TRUE))
  se_between <- var(coefs, na.rm = TRUE)
  best_subset_pooled_results[best_subset_pooled_results$Variable == var, "SE"] <- sqrt(se_within + (1 + 1/m
}

# filter non-zero mean coefficients and sort by mean
best_subset_selected_sorted <- best_subset_pooled_results %>%
  filter(Mean != 0, Variable != "(Intercept)") %>%
  arrange(desc(Mean))

# display results in a table
```

```r
best_sub_table <- kable(best_subset_selected_sorted,
                        caption = "Best Subset Model Selected Variables
                        (Non-Zero Coefficients)") %>%
  kable_styling(full_width = F, font_size = 12)

best_sub_table


best_subset_auc_values <- numeric(m)


plot(0, 0, type = "n", xlim = c(1, 0), ylim = c(0, 1),
     xlab = "1 - Specificity", ylab = "Sensitivity",
     main = "ROC Curves Across Imputed Datasets - Best Subset")
abline(a = 1, b = -1, col = "grey", lty = 2)

# loop through each imputed dataset to calculate and plot ROC curves
for (i in 1:m) {
  # convert predictions to numeric form if needed
  predictions_numeric <- as.numeric(best_subset_test_predictions[[i]])

  # calculate ROC curve
  roc_curve_best_subset <- roc(Y_test, predictions_numeric)
  best_subset_auc_values[i] <- auc(roc_curve_best_subset)  # store AUC for dataset

  # plot each ROC curve with "mako" colors
  plot(roc_curve_best_subset, col = viridis(m, option = "mako")[i],
       lwd = 2, add = TRUE)
}

# calculate mean and standard deviation of AUC across imputations
mean_auc_best_subset <- mean(best_subset_auc_values)
sd_auc_best_subset <- sd(best_subset_auc_values)


text(x = 1, y = 0.95,
     labels = paste("Mean AUC:", round(mean_auc_best_subset, 3), "\nSD:",
                    round(sd_auc_best_subset, 3)),
     adj = 0, col = "black", cex = 0.8)

legend("bottomright", legend = paste("Imputed Dataset", 1:m),
       col = viridis(m, option = "mako"), lwd = 2)


# prepare df for calibration by combining predictions and observed values
calibration_data_best_subset <- do.call(rbind, lapply(1:m, function(i) {
    data.frame(
        predicted = as.vector(best_subset_test_predictions[[i]]),
        observed = as.numeric(Y_test) - 1
    )
}))

# bin predictions for calibration
calibration_data_best_subset <- calibration_data_best_subset %>%
    mutate(bin = cut(predicted, breaks = 5)) %>%
    group_by(bin) %>%
```

```
    summarize(
        expected = mean(predicted),
        observed = mean(observed),
        se = ifelse(n() > 1, sqrt((observed * (1 - observed)) / n()), NA)
    )

# Create calibration plot
ggplot(calibration_data_best_subset, aes(x = expected, y = observed)) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  geom_smooth(method = "loess", color = "black", se = TRUE, fill = "gray") +
  geom_smooth(method = "lm", color = "blue", linetype = "dotted", se = FALSE) +
  labs(
    x = "Expected Proportion",
    y = "Observed Proportion",
    title = "Calibration Plot for Best Subset Model"
  ) +
  theme_minimal()
```