Predicting Student Enrollment with Classification Models: A Machine Learning Approach

Jay Chang

2025-08-19

Contents

Abstract

In this project, I analyze Learnova's prospect data to identify the key factors that influence whether a user enrolls in an online course. Using exploratory data analysis and visualization, I examine demographic, behavioral, and marketing-related variables such as age, occupation status, engagement time, and ad exposure. I then develop and evaluate classification models—including logistic regression, random forest, and boosted tree models—using metrics like precision, recall, F1-score, and ROC-AUC. The goal is to provide actionable insights into the drivers of enrollment and recommend strategies that help Learnova better target and convert prospective learners.

Business Context

The digital learning sector is rapidly growing, projected to hit $370B by 2026 with ~8.5% CAGR. Learnova is a startup targeting students and professionals with advanced tech courses. The challenge: not all leads convert. Outreach (calls, emails, ads) costs time and money — so prioritizing high-probability leads is essential.

How do we identify which prospects are most likely to enroll, so we can prioritize outreach and improve efficiency?

Data Overview

Dataset covers demographics like age and occupation, engagement behaviors such as profile completion, site visits, and time spent, as well as acquisition channels like ads, forums, and referrals. Our target variable is whether someone actually enrolled or not (enrollment_status). Before any further analysis, we have an initial consensus of 4 variable groups that could influence someone's decision to enroll: - Demographics: Age, Occupation - Behavior: Profile Status, Visits, Engagement Time, Pages per Session - Channel: Initial Contact, Recent Engagement - Advertising Source: Ads, Forums, Referrals

Inital Key Insights

- **Profile Completion:** The most important factor.
- **Engagement Depth:** The more engaged they are, the higher the chance of enrolling.
- **Source Quality:** Referrals likely convert better than ads.
- **Occupation:** Job seekers probably enroll more, followed by professionals, then students.
- **Initial Contact:** Mobile app users are often more engaged than website-only visitors.

Initial Data Analysis

Loading necessary libraries, reading in csv file

```
library(readr)
library(tidyverse)
library(lubridate)
library(scales)
library(broom)

df <- read_csv("C:/Users/jaych/Downloads/Learnova_Leads (1).csv")
```

Numeric Summaries

```
num_cols <- df %>%
  select(where(is.numeric)) %>%
  names()

num_summary <- df %>%
  select(all_of(num_cols)) %>%
```

```
  summary()
num_summary
```

```
##     user_age      site_visits      engagement_time   avg_pages_per_session
##  Min.   :18.0   Min.   : 0.000   Min.   :   0.0   Min.   : 0.000
##  1st Qu.:36.0   1st Qu.: 2.000   1st Qu.: 148.8   1st Qu.: 2.078
##  Median :51.0   Median : 3.000   Median : 376.0   Median : 2.792
##  Mean   :46.2   Mean   : 3.567   Mean   : 724.0   Mean   : 3.026
##  3rd Qu.:57.0   3rd Qu.: 5.000   3rd Qu.:1336.8   3rd Qu.: 3.756
##  Max.   :63.0   Max.   :30.000   Max.   :2537.0   Max.   :18.434
##  enrollment_status
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2986
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Class Balance and Overall Enrollment Rate

```
class_balance <- df %>%
  count(enrollment_status) %>%
  mutate(proportion = n / sum(n))

class_balance
```

```
## # A tibble: 2 x 3
##   enrollment_status     n proportion
##               <dbl> <int>      <dbl>
## 1                 0  3235      0.701
## 2                 1  1377      0.299
```

Enrollment Rates

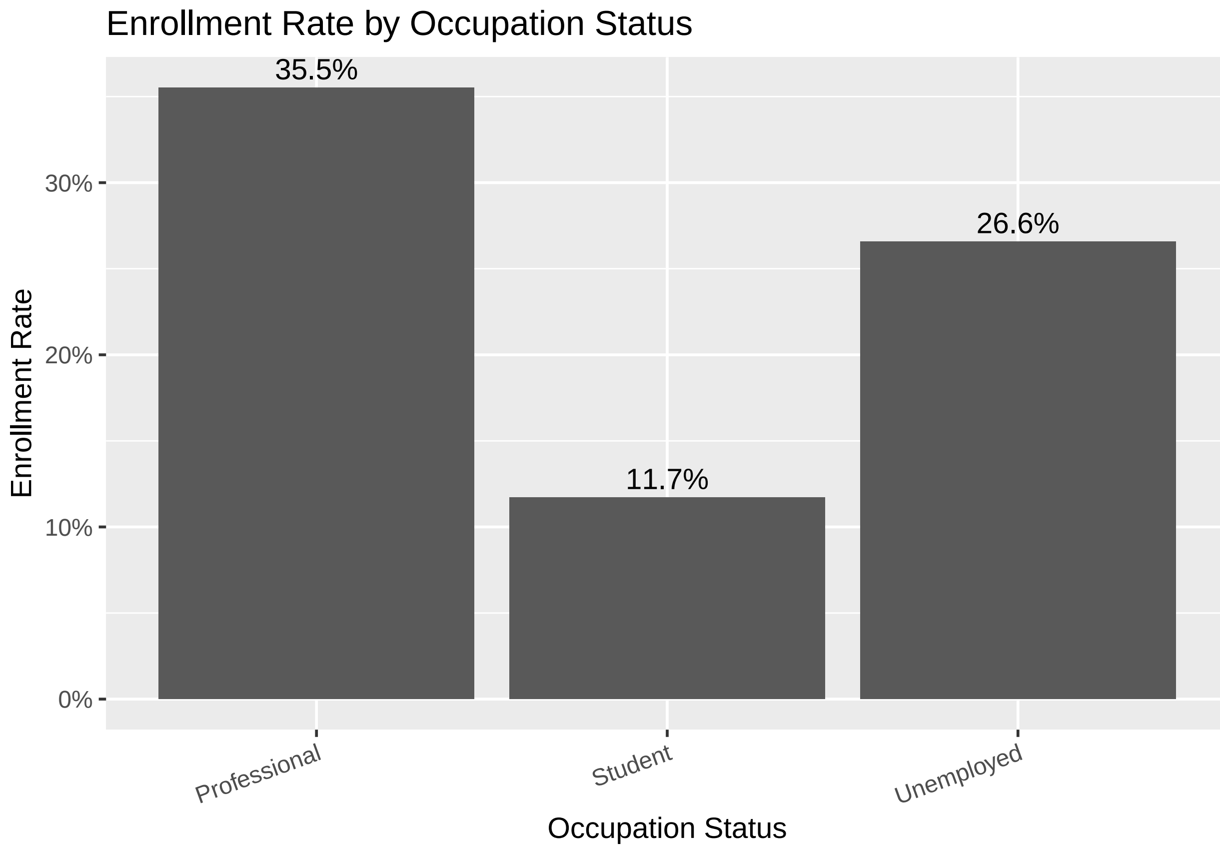
Enrollment Rate by Key Categorical Variables

```
rate_by <- function(data, col) {
  data %>%
    group_by({{ col }}) %>%
    summarise(
      n = n(),
      enroll_rate = mean(enrollment_status == 1)
    ) %>%
    arrange(desc(enroll_rate))
}

by_occupation <- rate_by(df, occupation_status)
by_profile <- rate_by(df, profile_status)
by_initial <- rate_by(df, initial_contact)
```

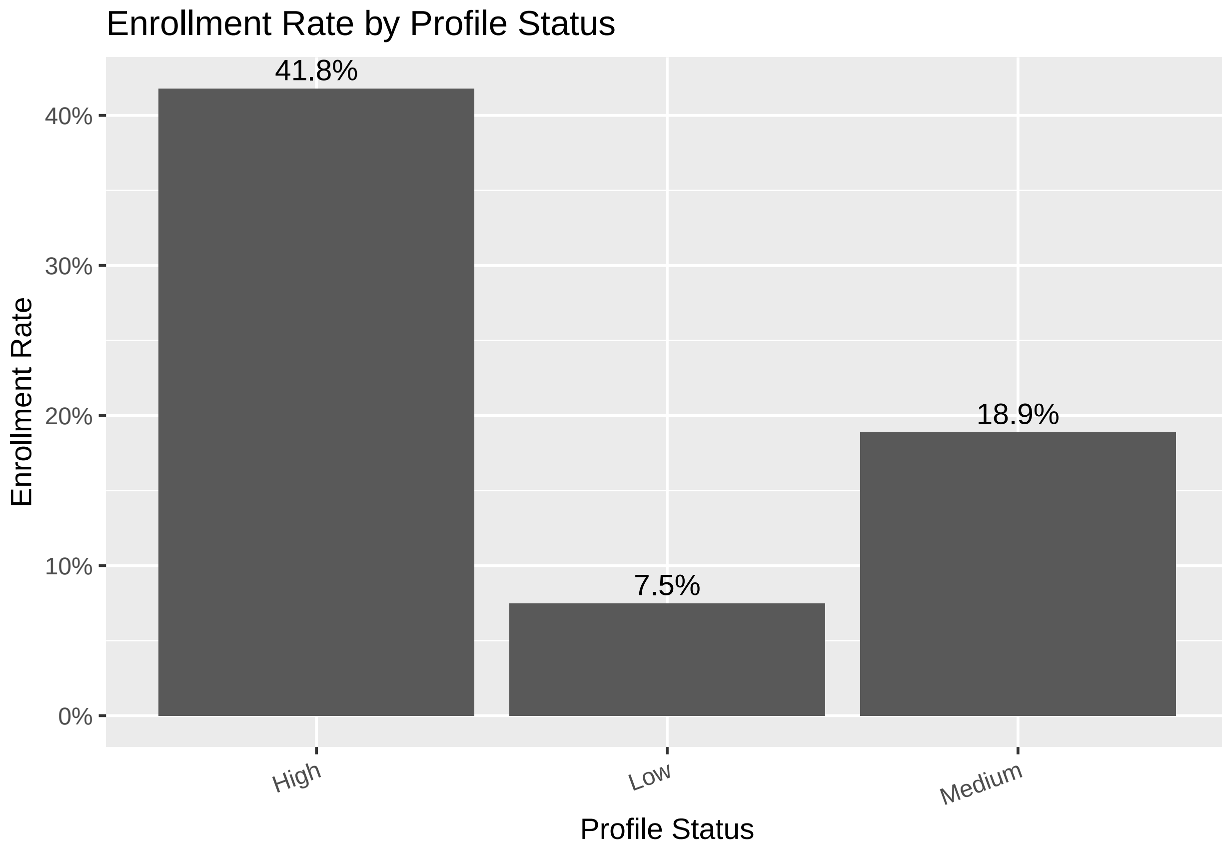
```r
by_recent <- rate_by(df, recent_engagement)

by_occupation; by_profile; by_initial; by_recent
```

```
## # A tibble: 3 x 3
##   occupation_status     n enroll_rate
##   <chr>             <int>       <dbl>
## 1 Professional       2616       0.355
## 2 Unemployed         1441       0.266
## 3 Student             555       0.117

## # A tibble: 3 x 3
##   profile_status     n enroll_rate
##   <chr>          <int>       <dbl>
## 1 High            2264      0.418
## 2 Medium          2241      0.189
## 3 Low              107      0.0748

## # A tibble: 2 x 3
##   initial_contact     n enroll_rate
##   <chr>           <int>       <dbl>
## 1 Website          2542       0.456
## 2 Mobile App       2070       0.105

## # A tibble: 3 x 3
##   recent_engagement     n enroll_rate
##   <chr>             <int>       <dbl>
## 1 Website Activity   1100       0.385
## 2 Email Activity     2278       0.303
## 3 Phone Activity     1234       0.213
```

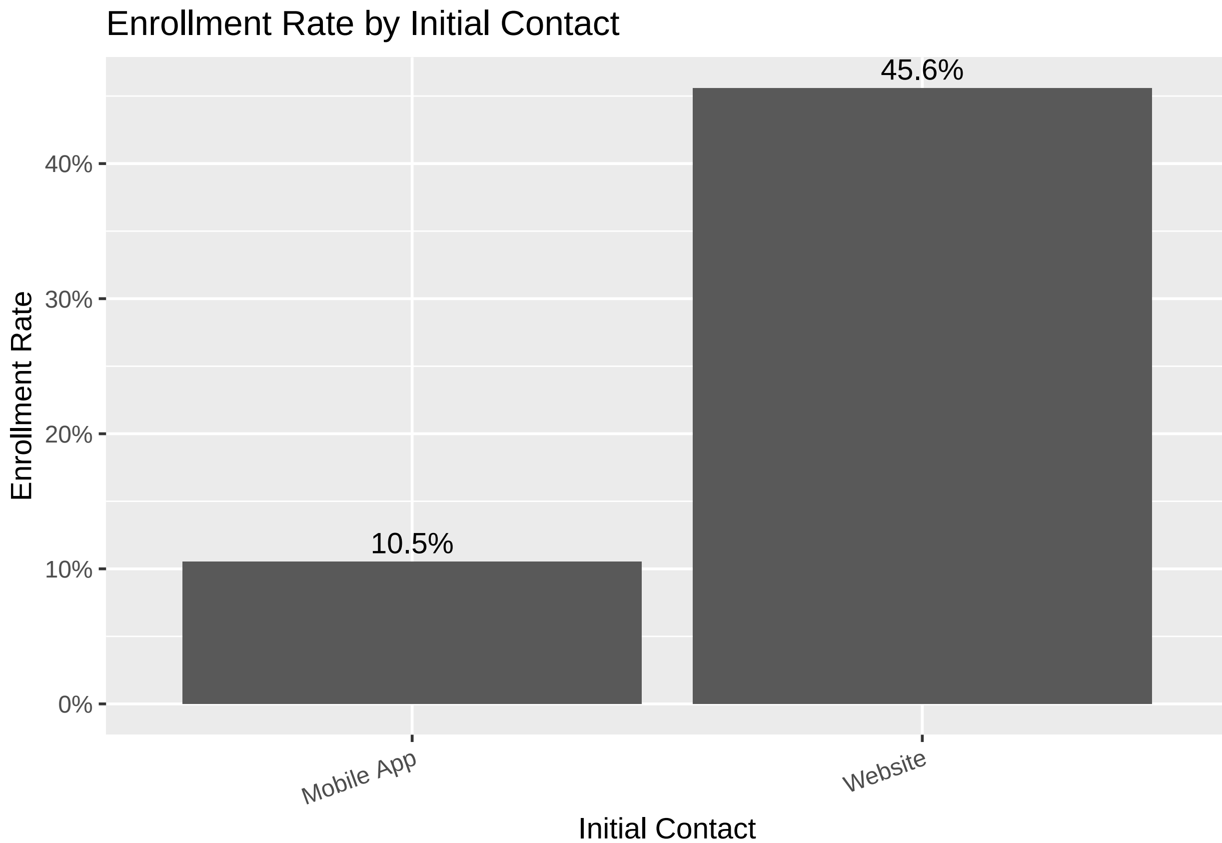
```r
plot_rate <- function(tbl, x, xlab) {
  ggplot(tbl, aes(x = {{ x }}, y = enroll_rate)) +
    geom_col() +
    geom_text(aes(label = percent(enroll_rate, accuracy = 0.1)), vjust = -0.4) +
    scale_y_continuous(labels = percent) +
    labs(title = paste("Enrollment Rate by", xlab), x = xlab, y = "Enrollment Rate") +
    theme(axis.text.x = element_text(angle = 20, hjust = 1))
}

plot_rate(by_occupation, occupation_status, "Occupation Status")
```



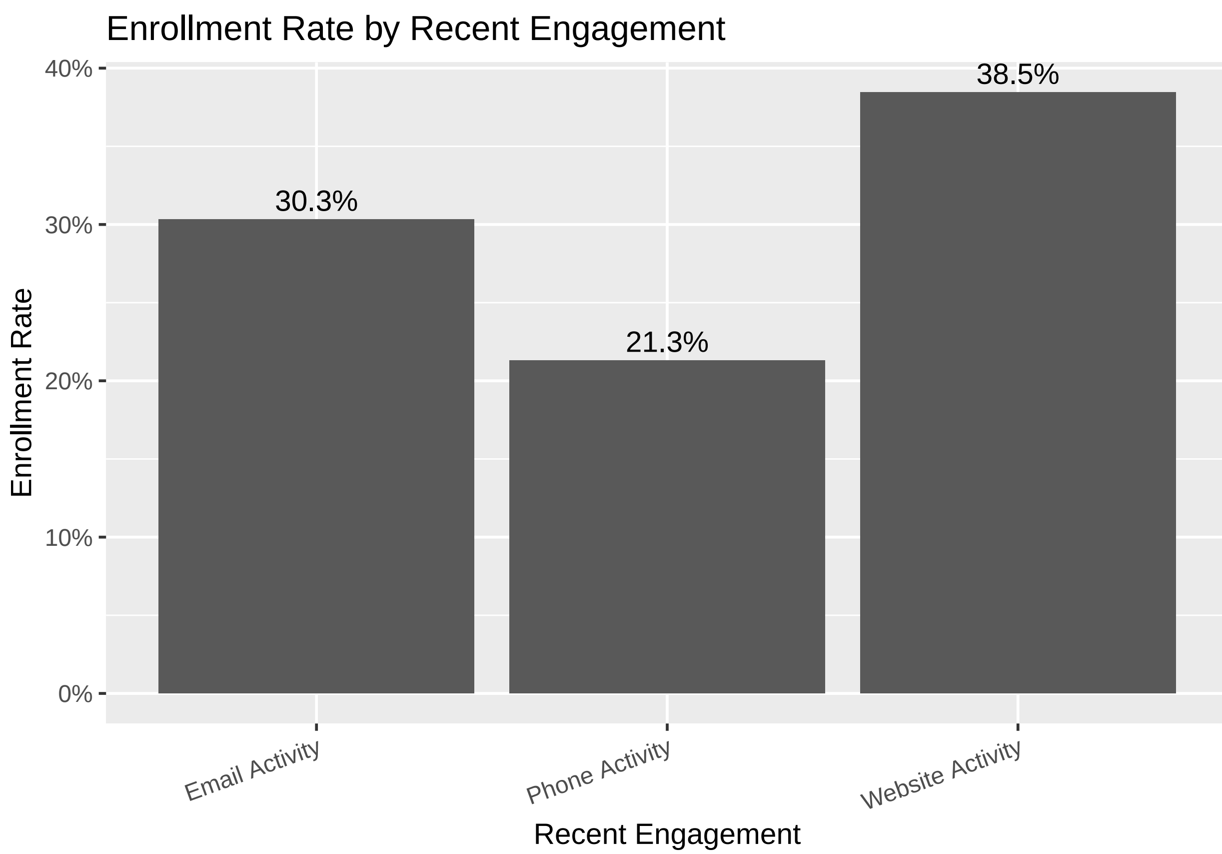

```r
plot_rate(by_profile, profile_status, "Profile Status")
```




```
plot_rate(by_initial, initial_contact, "Initial Contact")
```




```r
plot_rate(by_recent, recent_engagement, "Recent Engagement")
```



Enrollment Rate by Advertisement Source

```
source_cols <- c("newspaper_ad","magazine_ad","online_ad","edu_forums","word_of_mouth")

source_rates <- map_dfr(source_cols, function(col) {
  df %>%
    mutate(yes_no = .data[[col]]) %>%
    group_by(yes_no) %>%
    summarise(
      n = n(),
      enroll_rate = mean(enrollment_status == 1)
    ) %>%
    mutate(ad_source = col)
}) %>%
  relocate(ad_source)

source_rates
```

```
## # A tibble: 10 x 4
##    ad_source     yes_no     n enroll_rate
##    <chr>         <chr>  <int>       <dbl>
##  1 newspaper_ad  No      4115       0.296
##  2 newspaper_ad  Yes      497       0.320
##  3 magazine_ad   No      4379       0.297
##  4 magazine_ad   Yes      233       0.322
##  5 online_ad     No      4085       0.296
```

```
##   6 online_ad     Yes       527        0.319
##   7 edu_forums    No       3907        0.302
##   8 edu_forums    Yes       705        0.279
##   9 word_of_mouth No       4519        0.291
##  10 word_of_mouth Yes        93        0.677
```

Enrollment Rate by Numeric Variables (Engagement)

```
iqr_bin <- function(x) {
  qs <- quantile(x, probs = c(.25, .75), na.rm = TRUE)
  cut(x,
      breaks = c(-Inf, qs[1], qs[2], Inf),
      labels = c("Low", "Medium", "High"),
      include.lowest = TRUE, right = TRUE, ordered_result = TRUE)
}

df <- df %>%
  mutate(
    site_visits_bin           = iqr_bin(site_visits),
    engagement_time_bin       = iqr_bin(engagement_time),
    avg_pages_per_session_bin = iqr_bin(avg_pages_per_session),
    user_age_bin              = iqr_bin(user_age)
  )

rate_by_name <- function(data, var) {
  data %>%
    group_by(.data[[var]]) %>%
    summarise(
      n = n(),
      enroll_rate = mean(enrollment_status == 1),
      .groups = "drop"
    ) %>%
    arrange(desc(enroll_rate))
}

by_user_age <- rate_by_name(df, "user_age_bin")
by_site_visits <- rate_by_name(df, "site_visits_bin")
by_engage_time <- rate_by_name(df, "engagement_time_bin")
by_avg_pages <- rate_by_name(df, "avg_pages_per_session_bin")

by_user_age; by_site_visits; by_engage_time; by_avg_pages
```

```
## # A tibble: 3 x 3
##   user_age_bin     n enroll_rate
##   <ord>        <int>       <dbl>
## 1 Medium        2324       0.331
## 2 High          1081       0.322
## 3 Low           1207       0.215
```

```
## # A tibble: 3 x 3
##   site_visits_bin     n enroll_rate
##   <ord>           <int>       <dbl>
## 1 Medium           1557       0.305
```

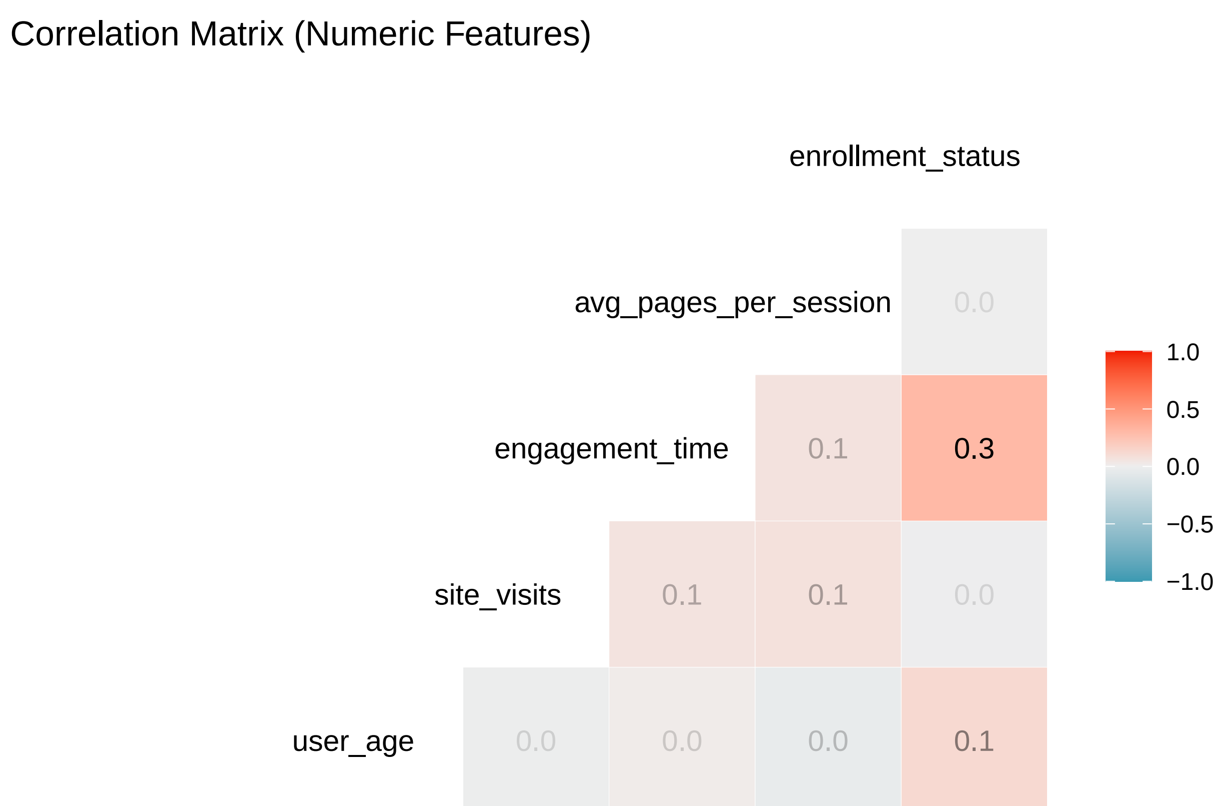
```
## 2 Low              2158       0.297
## 3 High              897       0.292
```

```
## # A tibble: 3 x 3
##   engagement_time_bin     n enroll_rate
##   <ord>                <int>       <dbl>
## 1 High                  1153       0.497
## 2 Medium                2306       0.273
## 3 Low                   1153       0.152
```

```
## # A tibble: 3 x 3
##   avg_pages_per_session_bin     n enroll_rate
##   <ord>                      <int>       <dbl>
## 1 Medium                      2306       0.310
## 2 Low                         1153       0.293
## 3 High                        1153       0.282
```

Correlation Heatmap (Numeric Variables)

```
df_num <- df %>% select(all_of(num_cols))
# Quick correlation heatmap (removes columns with 0 variance)
df_num_nzv <- df_num %>% select(where(~ sd(.x, na.rm = TRUE) > 0))
GGally::ggcorr(df_num_nzv, label = TRUE, label_alpha = TRUE, hjust = 0.8, layout.exp = 2) +
  ggtitle("Correlation Matrix (Numeric Features)")
```



Assessing Classification Modeling

1. Logistic Regression
2. Decision Trees/Random Forest
3. Boosted Trees (XGBoost)

Performance Metrics

- Precision: Ensures outreach focuses on true high-probability prospects
- Recall: Missing potential customers
- F1-Score: Balance between precision and recall
- ROC-AUC: Overall ranking ability

Actionable Insights and Strategic Recommendations

Refined Insights

Profile Completion, Engagement Depth, and Source Quality remains driving factors of learner enrollment.

- Occupation Status: Professionals > Unemployed/Job Seekers > Students
- Initial Contact: Website > Mobile App
- Most Recent Engagement: Website Activity > Email > Phone

Recommendations

1. Prioritize Hot Leads - High profile completion - High site engagement - Referral-based leads

2. Deploy Predictive Scoring - "Probability of Enrolling" score - Ranked leads and efficient outreach

3. Outreach Sequencing - Website –> Email –> Phone - Occupation-based marketing, tailored campaigns

4. Referral Programs - Discounts for friends

5. Pilot Test - 2-Week Plan - Utilize predictive score, guide outreach, comparing enrollment lift, iterate on key features