Term Deposit Prediction & Marketing Analysis

Targeting Customers that are likely to opt into a Term Deposit service

Uche Kalu

4/9/2022

The goal of this analysis is to determine which features in the bank data relates to the bank's marketing campaign goal of having customers opt-into a TERM DEPOSIT (financial product)

Executive Summary

- Duration since last contact and success of prior enrollments are correlated to success of Term Deposit opt-in

- A strategy targeting customers that haven't been contacted in 319 days and have those with prior enrollments yields 32% positive outcomes (enrollments) vs 4.3% otherwise.

Analyzing and Investigating the Data

```
path   <- "bank_term_deposit_marketing_analysis.xlsx"
sheets <- excel_sheets(path)  # displays all d sheet names in d workbook
sheets
```

```
##  [1] "PROCEDURE"                    "DATA DESCRIPTION"
##  [3] "Step 1 - Collect Information" "CLIENT_INFO"
##  [5] "LOAN_HISTORY"                 "MARKETING HISTORY"
##  [7] "SUBSCRIPTION HISTORY"         "Step 2 - Merge Information"
##  [9] "CLIENT_MERGE"                 "Step 3 - Marketing Analysis"
## [11] "DAILY RANGE"                  "JOB ANALYSIS"
```

Our dataset parameters are in sheets 4 to 7 so we select only 4th to 7th sheets and combine all their data into one sheet.

```
data_joined_tbl <- sheets[4:7] %>%
  map(~ read_excel(path = path, sheet = .)) %>%
  reduce(left_join)
```

```
## Joining, by = "ID"
## Joining, by = "ID"
## Joining, by = "ID"
```

```
data_joined_tbl %>% glimpse()
```

```
## Rows: 45,211
## Columns: 18
## $ ID           <chr> "2836", "2837", "2838", "2839", "2840", "2841", "2842", "~
## $ AGE          <dbl> 58, 44, 33, 47, 33, 35, 28, 42, 58, 43, 41, 29, 53, 58, 5~
## $ JOB          <chr> "management", "technician", "entrepreneur", "blue-collar"~
## $ MARITAL      <chr> "married", "single", "married", "married", "single", "mar~
## $ EDUCATION    <chr> "tertiary", "secondary", "secondary", "unknown", "unknown~
## $ DEFAULT      <chr> "no", "no", "no", "no", "no", "no", "no", "yes", "no", "n~
## $ BALANCE      <dbl> 2143, 29, 2, 1506, 1, 231, 447, 2, 121, 593, 270, 390, 6,~
## $ HOUSING      <chr> "yes", "yes", "yes", "yes", "no", "yes", "yes", "yes", "y~
## $ LOAN         <chr> "no", "no", "yes", "no", "no", "no", "yes", "no", "no", "~
## $ CONTACT      <chr> "unknown", "unknown", "unknown", "unknown", "unknown", "u~
## $ DAY          <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
## $ MONTH        <chr> "may", "may", "may", "may", "may", "may", "may", "may", "~
## $ DURATION     <dbl> 261, 151, 76, 92, 198, 139, 217, 380, 50, 55, 222, 137, 5~
## $ CAMPAIGN     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ PDAYS        <dbl> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -~
## $ PREVIOUS     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ POUTCOME     <chr> "unknown", "unknown", "unknown", "unknown", "unknown", "u~
## $ TERM_DEPOSIT <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no~
```

Analyze and Prepare the Data using recipe package

Modeling and Machine learning often involves a response feature (Enrolled in TERM_DEPOSIT, yes/no) and many predictor features (AGE, JOB, MARITAL, etc). We are going to determine which predictors are related to the response variable. We do this using Binary Correlation Analysis

```
recipe_obj <- recipe(~ ., data = data_joined_tbl) %>%
  step_rm(ID) %>%
  step_discretize(all_numeric(), options = list(min_unique = 1)) %>%
  step_dummy(all_nominal(), one_hot = TRUE, naming = partial(dummy_names, sep = "__")) %>%
  prep()

data_transformed_tbl <- data_joined_tbl %>%
  bake(recipe_obj, new_data = .)
```

```
data_transformed_tbl %>% glimpse()
```

```
## Rows: 45,211
## Columns: 74
## $ AGE__bin_missing      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ AGE__bin1             <dbl> 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, ~
## $ AGE__bin2             <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ AGE__bin3             <dbl> 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, ~
## $ AGE__bin4             <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, ~
## $ JOB__admin.           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, ~
## $ JOB__blue.collar      <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__entrepreneur     <dbl> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__housemaid        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__management       <dbl> 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__retired          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, ~
```

```
## $ JOB__self.employed    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__services         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ~
## $ JOB__student          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__technician       <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, ~
## $ JOB__unemployed       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ JOB__unknown          <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MARITAL__divorced     <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ MARITAL__married      <dbl> 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, ~
## $ MARITAL__single       <dbl> 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, ~
## $ EDUCATION__primary    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, ~
## $ EDUCATION__secondary  <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, ~
## $ EDUCATION__tertiary   <dbl> 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ EDUCATION__unknown    <dbl> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ~
## $ DEFAULT__no           <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ DEFAULT__yes          <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ BALANCE__bin_missing  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ BALANCE__bin1         <dbl> 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, ~
## $ BALANCE__bin2         <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, ~
## $ BALANCE__bin3         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ~
## $ BALANCE__bin4         <dbl> 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ HOUSING__no           <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ HOUSING__yes          <dbl> 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ LOAN__no              <dbl> 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ LOAN__yes             <dbl> 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CONTACT__cellular     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CONTACT__telephone    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CONTACT__unknown      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ DAY__bin_missing      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ DAY__bin1             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ DAY__bin2             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ DAY__bin3             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ DAY__bin4             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__apr            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__aug            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__dec            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__feb            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__jan            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__jul            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__jun            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__mar            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__may            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ MONTH__nov            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__oct            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ MONTH__sep            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ DURATION__bin_missing <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ DURATION__bin1        <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, ~
## $ DURATION__bin2        <dbl> 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
## $ DURATION__bin3        <dbl> 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ DURATION__bin4        <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, ~
## $ CAMPAIGN__bin_missing <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CAMPAIGN__bin1        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ CAMPAIGN__bin2        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ CAMPAIGN__bin3        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ PDAYS__bin_missing    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

```
## $ PDAYS__bin1           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ PREVIOUS__bin_missing <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ PREVIOUS__bin1        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ POUTCOME__failure     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ POUTCOME__other       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ POUTCOME__success     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ POUTCOME__unknown     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TERM_DEPOSIT__no      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TERM_DEPOSIT__yes     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Correlation of various features to Term Deposit enrollment.

We perform a correlation analysis between the response variable(TERM_DEPOSIT_yes) and the rest of the features.

```
correlation_tbl <- data_transformed_tbl %>%
  cor(y = data_transformed_tbl$TERM_DEPOSIT__yes) %>%
  as_tibble(rownames = "feature") %>%
  rename(TERM_DEPOSIT__yes = V1) %>%
  separate(feature, into = c("feature", "bin"), sep = "__") %>%
  filter(!is.na(TERM_DEPOSIT__yes)) %>%    # removes all NA values
  filter(!str_detect(feature, "TERM_DEP")) %>%
  arrange(abs(TERM_DEPOSIT__yes) %>% desc()) %>%
  mutate(feature = as_factor(feature) %>% fct_rev())

correlation_tbl
```

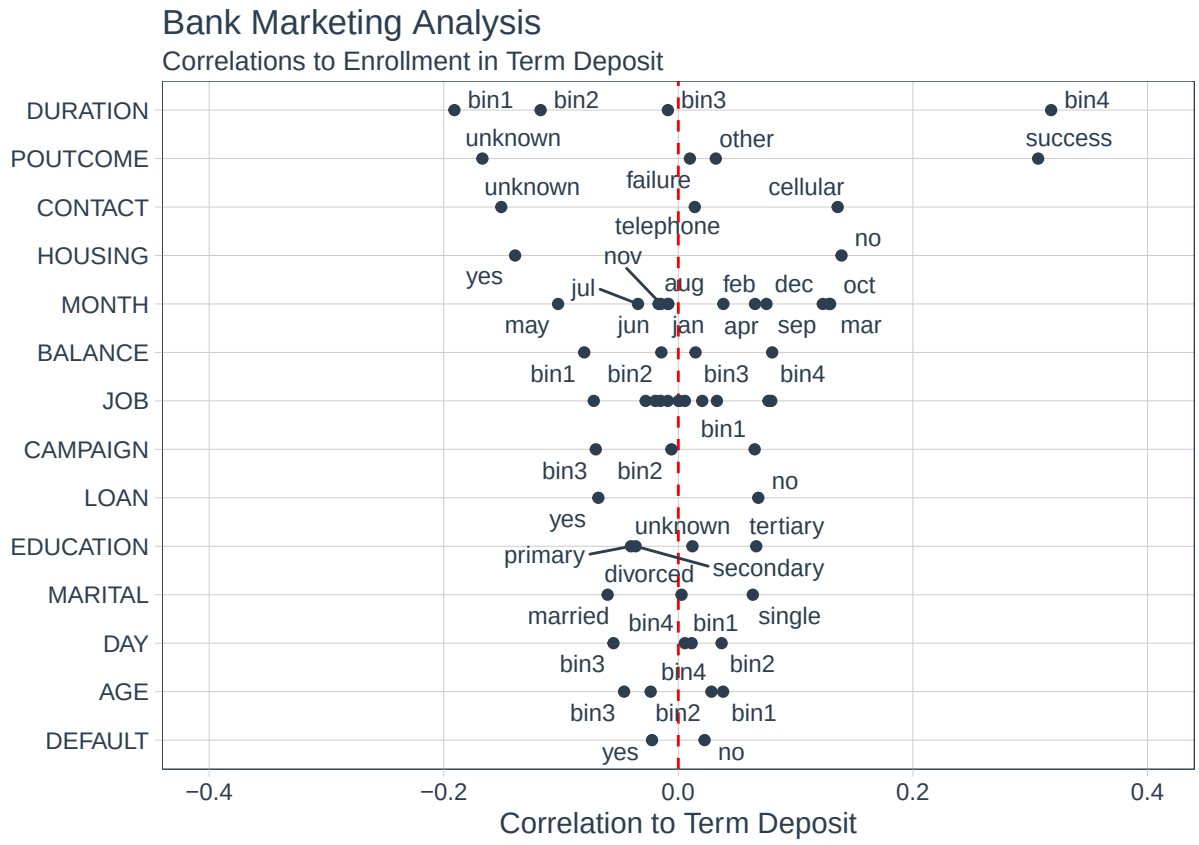
```
## # A tibble: 63 x 3
##    feature  bin      TERM_DEPOSIT__yes
##    <fct>    <chr>                <dbl>
##  1 DURATION bin4                 0.318
##  2 POUTCOME success              0.307
##  3 DURATION bin1                -0.191
##  4 POUTCOME unknown             -0.167
##  5 CONTACT  unknown             -0.151
##  6 HOUSING  no                   0.139
##  7 HOUSING  yes                 -0.139
##  8 CONTACT  cellular             0.136
##  9 MONTH    mar                  0.129
## 10 MONTH    oct                  0.129
## # ... with 53 more rows
```

Visualize the Correlations Funnel using Scatterplot

Examining the result chart, we deduce that Duration since last contact and success of prior enrollments are correlated to success of Term Deposit opt-in.

```
correlation_tbl %>%
  ggplot(aes(TERM_DEPOSIT__yes, y = feature, text = bin)) +
  geom_vline(xintercept = 0, linetype = 2, color = "red") +
  geom_point(color = "#2c3e50") +   # creates datapoints in d charts
  geom_text_repel(aes(label = bin), size = 3, color = "#2c3e50") +
```

```
  expand_limits(x = c(-0.4, 0.4)) +  # increasing d range of values in x-axis
  theme_tq() +
  labs(title = "Bank Marketing Analysis",
       subtitle = "Correlations to Enrollment in Term Deposit",
       y = "", x = "Correlation to Term Deposit")
```



Lets Interpret the Correlations

We have bins as values for Duration, Balance, Campaign, Day and Age in the chart and we need to know what values are in them bins?

```
recipe_obj %>% tidy()
```

```
## # A tibble: 3 x 6
##   number operation type       trained skip  id
##    <int> <chr>     <chr>      <lgl>   <lgl> <chr>
## 1      1 step      rm         TRUE    FALSE rm_iitTS
## 2      2 step      discretize TRUE    FALSE discretize_J6imL
## 3      3 step      dummy      TRUE    FALSE dummy_eAE3O
```

```
bins_tbl <- recipe_obj %>% tidy(2)
bins_tbl
```

```
## # A tibble: 28 x 3
##    terms   value id
##    <chr>   <dbl> <chr>
##  1 AGE      -Inf discretize_J6imL
##  2 AGE        33 discretize_J6imL
##  3 AGE        39 discretize_J6imL
##  4 AGE        48 discretize_J6imL
##  5 AGE       Inf discretize_J6imL
##  6 BALANCE  -Inf discretize_J6imL
##  7 BALANCE    72 discretize_J6imL
##  8 BALANCE   448 discretize_J6imL
##  9 BALANCE  1428 discretize_J6imL
## 10 BALANCE   Inf discretize_J6imL
## # ... with 18 more rows
```

```
View(bins_tbl)

bins_tbl %>% filter(terms == "DURATION")
```

```
## # A tibble: 5 x 3
##   terms    value id
##   <chr>    <dbl> <chr>
## 1 DURATION  -Inf discretize_J6imL
## 2 DURATION   103 discretize_J6imL
## 3 DURATION   180 discretize_J6imL
## 4 DURATION   319 discretize_J6imL
## 5 DURATION   Inf discretize_J6imL
```

STRATEGY

A strategy targeting customers that haven't been contacted in 319 days and have those with prior enrollments yields 32% positive outcomes (enrollments) vs 4.3% otherwise.

We focus on value - BIN4 in DURATION column and value - success in POUTCOME column since they are the top columns that affects Term Deposit enrollment and their stated values are the only values at far RHS

```
strategy_tbl <- data_joined_tbl %>%
  select(DURATION, POUTCOME, TERM_DEPOSIT) %>%
  mutate(POTENTIAL = case_when(DURATION > 319 ~ "High Potential",
                               POUTCOME == "success" ~ "High Potential",
                               TRUE ~ "Normal")) %>%
  group_by(POTENTIAL) %>%  # displays unique values in d stated column
  count(TERM_DEPOSIT) %>%
  mutate(prop = n / sum(n)) %>%
  ungroup() %>%
  mutate(label_text = str_glue("n: {n}
                               prop: {scales::percent(prop)}"))
```

```
strategy_tbl
```

```
## # A tibble: 4 x 5
```

```
##   POTENTIAL      TERM_DEPOSIT     n   prop label_text
##   <chr>          <chr>        <int>  <dbl> <glue>
## 1 High Potential no            8364 0.683  n: 8364
## prop: 68%
## 2 High Potential yes           3887 0.317  n: 3887
## prop: 32%
## 3 Normal         no           31558 0.957  n: 31558
## prop: 96%
## 4 Normal         yes           1402 0.0425 n: 1402
## prop: 4%
```

REPORT RESULTS using a stacked Bar chart

```
strategy_tbl %>%
  ggplot(aes(POTENTIAL, prop, fill = TERM_DEPOSIT)) +
  geom_col() +
  geom_label(aes(label = label_text), fill = "white", color = "#2c3e50") +
  scale_fill_tq() +                       # changes d color of d stacked bars
  scale_y_continuous(labels = scales::percent_format()) +
  theme_tq() +
  labs(title = "Bank Marketing Strategy",
       subtitle = str_glue("Targeting customers that haven't been contacted in 319 days or those with p
```