Analyst Case Study

2025-01-27

```r
library(readxl)

file_path <- "/Users/sayanandrews/Downloads/Analyst case study dataset 1.xls"

# Read the Excel file
data <- read_excel(file_path)

head(data)
```

```
## # A tibble: 6 x 24
##   LeadCreated         FirstName Email         VendorLeadID CallStatus WidgetName
##   <dttm>              <chr>     <chr>         <chr>        <chr>      <chr>
## 1 2009-07-01 01:24:29 Dorinda   kanani@sandw~ FDF81FDA-A6~ <NA>       w-302252--~
## 2 2009-04-13 15:27:35 Presetta  clerk2@ustco~ 4190ACB7-50~ <NA>       w-300250--~
## 3 2009-04-21 07:22:59 Gina      wagoner_gina~ hFg80jf_R0C~ Unable to~ w-300250--~
## 4 2009-08-03 19:39:02 Kari      usa4ley@yaho~ jB01QgYZxkW~ Contacted~ w-302252--~
## 5 2009-04-13 10:18:37 Stephanie sr1lambert@e~ D5B32074-45~ <NA>       w-300250--~
## 6 2009-07-29 14:50:06 diana     diana.powers~ 94AB0C2C-0D~ <NA>       w-302252--~
## # i 18 more variables: PublisherZoneName <chr>, PublisherCampaignName <chr>,
## #   AddressScore <chr>, PhoneScore <chr>, AdvertiserCampaignName <chr>,
## #   State <chr>, DebtLevel <chr>, ‘IP Address‘ <lgl>, Partner <chr>,
## #   ReferralDomain <chr>, MarketingCampaign <chr>, AdGroup <chr>,
## #   Keyword <chr>, SearchQuery <chr>, ReferralURL <chr>,
## #   ‘ReferralURL Parameters‘ <chr>, LandingPageURL <chr>,
## #   ‘Landing Page URL Parameters‘ <chr>
```

Question 1 – Are we seeing any lead quality trends over time (improving, declining)? Are they statistically significant?

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

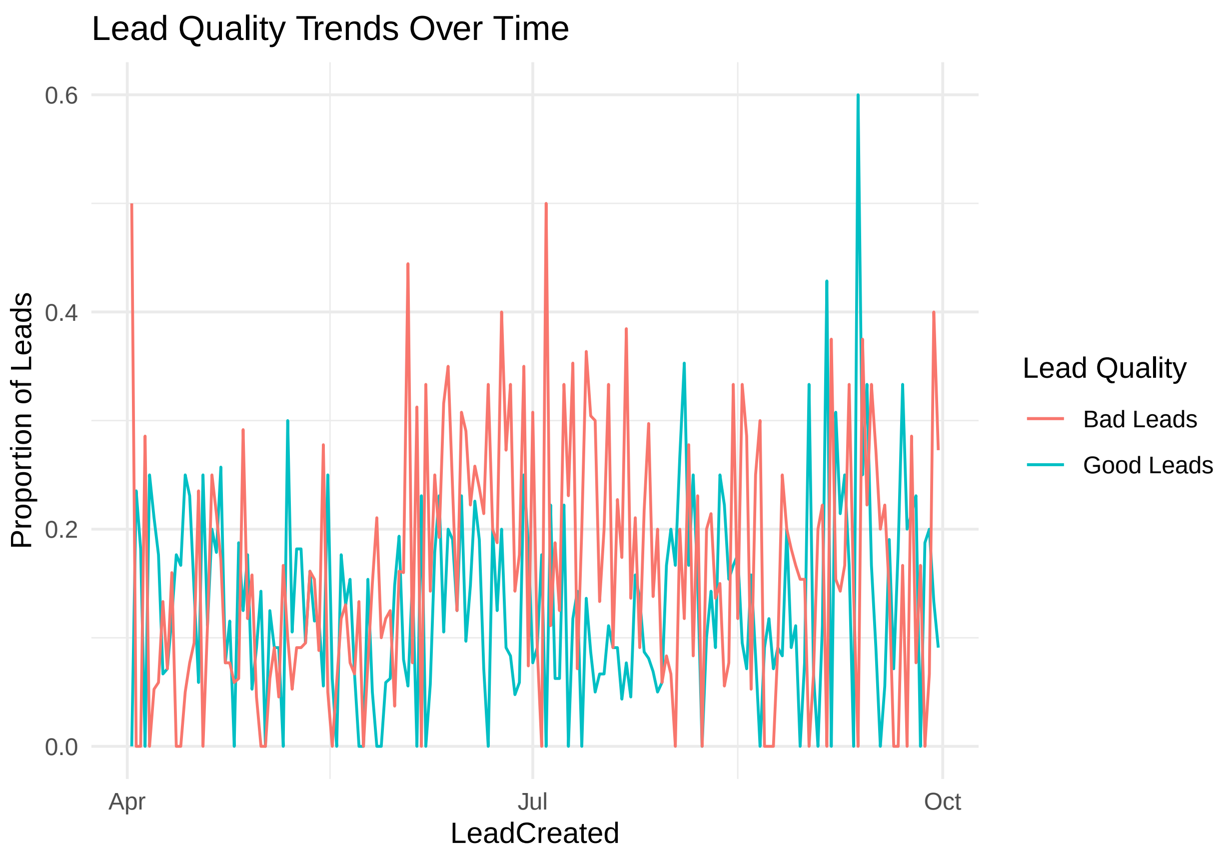
```r
library(ggplot2)

# Convert Lead Created Date to Date format
data$LeadCreated <- as.Date(data$LeadCreated)

# Define lead quality categories
good_leads <- c("Closed", "EP Sent", "EP Received", "EP Confirmed")
bad_leads <- c("Unable to contact - Bad Contact Information", "Contacted - Invalid Profile", "Contacted

# Calculate lead quality over time
lead_quality_time <- data %>%
  mutate(Quality = case_when(
    CallStatus %in% good_leads ~ "Good",
    CallStatus %in% bad_leads ~ "Bad",
    TRUE ~ "Unknown"
  )) %>%
  group_by(LeadCreated) %>%
  summarize(
    TotalLeads = n(),
    GoodLeads = sum(Quality == "Good"),
    BadLeads = sum(Quality == "Bad"),
    UnknownLeads = sum(Quality == "Unknown"),
    GoodLeadRate = GoodLeads / TotalLeads,
    BadLeadRate = BadLeads / TotalLeads
  )

# Plot lead quality trends over time
ggplot(lead_quality_time, aes(x = LeadCreated)) +
  geom_line(aes(y = GoodLeadRate, color = "Good Leads")) +
  geom_line(aes(y = BadLeadRate, color = "Bad Leads")) +
  labs(title = "Lead Quality Trends Over Time", y = "Proportion of Leads", color = "Lead Quality") +
  theme_minimal()
```



Based on the lead quality trends over time shown in the provided graph, I would conclude the following:

Lead Quality Trends:

1. **Good Leads**:
    - The proportion of good leads shows fluctuations over the observed period, with some peaks and troughs.
    - There is no clear upward or downward trend that suggests a consistent improvement or decline in lead quality. Instead, it varies around a relatively stable average.
2. **Bad Leads**:
    - Similarly, the proportion of bad leads also fluctuates over the same period.
    - The trend does not indicate a consistent pattern of increasing or decreasing bad lead proportions, instead showing variability around a stable average.

Statistical Significance:

- To determine if these trends are statistically significant, further statistical analysis is needed (e.g., time series analysis, hypothesis testing). Simply observing the graph suggests that there is variability, but it does not provide strong evidence of a significant trend in either direction.

Recommendation:

- **Monitor Ongoing Trends**: Continue to track lead quality over time to identify any emerging trends more conclusively.

- **Further Analysis**: Conduct additional statistical tests to assess if any observed fluctuations are statistically significant.
- **Actionable Insights**: Use this analysis in combination with other insights (such as drivers of lead quality) to inform strategic decisions on how to improve lead quality.

Question 2 – What can we learn about the drivers of "lead quality" from this dataset?

```r
# Convert CallStatus to a binary outcome (1 for good lead, 0 for bad lead)
data <- data %>%
  mutate(GoodLead = ifelse(CallStatus %in% good_leads, 1, 0))

# Logistic regression model
model <- glm(GoodLead ~ WidgetName + PublisherZoneName + PublisherCampaignName + AddressScore + PhoneSc

# Summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = GoodLead ~ WidgetName + PublisherZoneName + PublisherCampaignName +
##     AddressScore + PhoneScore, family = binomial, data = data)
##
## Coefficients: (1 not defined because of singularities)
##                                                         Estimate Std. Error
## (Intercept)                                            -12.34890  324.74390
## WidgetNamew-300250-DebtReduction1-1DC-BlueMeter         -0.02165    0.38670
## WidgetNamew-300250-DebtReduction1-1DC-CreditSolutions    0.51518    0.38356
## WidgetNamew-300250-DebtReduction1-1DC-Head2              0.01955    0.39095
## WidgetNamew-300250-DebtReduction1-1DC-Head3             -0.94063    0.53195
## WidgetNamew-300250-DebtReduction1-1DC-white              0.09291    1.12899
## WidgetNamew-300250-DebtReduction1-2DC-BlueMeter         -0.13136    0.41177
## WidgetNamew-300250-DebtReduction1-2DC-CreditSolutions   -0.42291    0.46318
## WidgetNamew-302252-DebtReduction1-1DC                   -0.32652    0.33285
## WidgetNamew-302252-DebtReduction1-1DC-CreditSolutions   -0.24502    0.29728
## WidgetNamew-302252-DebtReduction1-1DC-white             -0.47349    0.32161
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow        0.52965    0.45183
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow-blue  -0.17738    0.35050
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow-dark  -0.36615    0.40900
## PublisherZoneNameTopLeft-302252                         -0.09427    0.31440
## PublisherCampaignNameDebtReductionInc                         NA         NA
## AddressScore2                                            0.47304    0.51282
## AddressScore3                                            0.05380    0.46176
## AddressScore4                                           -0.90535    0.78142
## AddressScore5                                            0.12212    0.30670
## AddressScoreNULL                                         0.01798    0.35703
## PhoneScore2                                             10.35049  324.74382
## PhoneScore3                                             10.54448  324.74375
## PhoneScore4                                             10.05244  324.74385
## PhoneScore5                                             10.92020  324.74373
## PhoneScoreNULL                                          10.72284  324.74380
##                                                        z value Pr(>|z|)
```

```
## (Intercept)                                              -0.038    0.970
## WidgetNamew-300250-DebtReduction1-1DC-BlueMeter          -0.056    0.955
## WidgetNamew-300250-DebtReduction1-1DC-CreditSolutions     1.343    0.179
## WidgetNamew-300250-DebtReduction1-1DC-Head2               0.050    0.960
## WidgetNamew-300250-DebtReduction1-1DC-Head3              -1.768    0.077 .
## WidgetNamew-300250-DebtReduction1-1DC-white               0.082    0.934
## WidgetNamew-300250-DebtReduction1-2DC-BlueMeter          -0.319    0.750
## WidgetNamew-300250-DebtReduction1-2DC-CreditSolutions    -0.913    0.361
## WidgetNamew-302252-DebtReduction1-1DC                    -0.981    0.327
## WidgetNamew-302252-DebtReduction1-1DC-CreditSolutions    -0.824    0.410
## WidgetNamew-302252-DebtReduction1-1DC-white              -1.472    0.141
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow         1.172    0.241
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow-blue   -0.506    0.613
## WidgetNamew-302252-DebtReduction1-1DC-yellowarrow-dark   -0.895    0.371
## PublisherZoneNameTopLeft-302252                          -0.300    0.764
## PublisherCampaignNameDebtReductionInc                        NA       NA
## AddressScore2                                             0.922    0.356
## AddressScore3                                             0.117    0.907
## AddressScore4                                            -1.159    0.247
## AddressScore5                                             0.398    0.691
## AddressScoreNULL                                          0.050    0.960
## PhoneScore2                                               0.032    0.975
## PhoneScore3                                               0.032    0.974
## PhoneScore4                                               0.031    0.975
## PhoneScore5                                               0.034    0.973
## PhoneScoreNULL                                            0.033    0.974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2335.6  on 3020  degrees of freedom
## Residual deviance: 2294.7  on 2996  degrees of freedom
## AIC: 2344.7
##
## Number of Fisher Scoring iterations: 11
```

Based on the output of the logistic regression model, here are some insights into the drivers of "lead quality":

Key Drivers of Lead Quality:

1. **WidgetName**:
   - Some widget names have different impacts on lead quality. For instance:
     - `WidgetNamew-300250-DebtReduction1-1DC-CreditSolutions` has a positive coefficient (0.51518), suggesting it is associated with higher lead quality.
     - `WidgetNamew-300250-DebtReduction1-1DC-Head3` has a negative coefficient (-0.94063), indicating it is associated with lower lead quality.
   - However, many widget names have non-significant p-values, indicating their impact on lead quality is not statistically significant.
2. **PublisherZoneName**:
   - `PublisherZoneNameTopLeft-302252` has a negative coefficient (-0.09427) but a high p-value (0.764), suggesting its impact on lead quality is not statistically significant.

3. **PublisherCampaignName**:

   - The effect of `PublisherCampaignNameDebtReductionInc` is not defined due to singularities, which means it has been excluded from the model.

4. **AddressScore**:

   - `AddressScore2` has a positive coefficient (0.47304) but a non-significant p-value (0.356).
   - `AddressScore4` has a negative coefficient (-0.90535) with a p-value of 0.247, indicating a potentially negative impact on lead quality, but it is not statistically significant.

5. **PhoneScore**:

   - PhoneScore variables have very high coefficients but also very high standard errors, leading to non-significant p-values. This suggests some issues with the data, potentially collinearity or small sample size for these scores.

Summary:

- **WidgetName**: There are specific widget names that seem to have a more significant impact on lead quality. For instance, `WidgetNamew-300250-DebtReduction1-1DC-CreditSolutions` has a positive association with lead quality.
- **PublisherZoneName** and **PublisherCampaignName**: These do not show significant effects on lead quality based on the model output.
- **AddressScore** and **PhoneScore**: The significance of AddressScore and PhoneScore in predicting lead quality is not strongly supported by the model, likely due to collinearity or insufficient data for certain scores.

Recommendations:

1. **Focus on Effective Widgets**: Given the positive impact of certain widget names, consider using those more prominently or testing variations to improve lead quality.
2. **Further Data Examination**: Investigate potential collinearity or data issues, particularly with PhoneScore, to ensure more reliable results.
3. **Segment Analysis**: Perform additional segment analysis to identify other potential drivers of lead quality that may not have been captured in this model.

Question 3 – If the advertiser says they will increase our CPL by 20% (i.e., $30 to $33) if we increase our lead quality by 20% (i.e., from 8.0% to 9.6%), do we see any opportunities to do that here? What kinds of things could we do?

```r
# Analyze lead quality by WidgetName
widget_quality <- data %>%
  group_by(WidgetName) %>%
  summarize(
    TotalLeads = n(),
    GoodLeads = sum(GoodLead),
    GoodLeadRate = GoodLeads / TotalLeads
  ) %>%
  arrange(desc(GoodLeadRate))

# Analyze lead quality by PublisherCampaignName
campaign_quality <- data %>%
```

```r
  group_by(PublisherCampaignName) %>%
  summarize(
    TotalLeads = n(),
    GoodLeads = sum(GoodLead),
    GoodLeadRate = GoodLeads / TotalLeads
  ) %>%
  arrange(desc(GoodLeadRate))

# Display top WidgetNames and PublisherCampaignNames by lead quality
head(widget_quality, 10)
```

```
## # A tibble: 10 x 4
##    WidgetName                                  TotalLeads GoodLeads GoodLeadRate
##    <chr>                                            <int>     <dbl>        <dbl>
##  1 w-302252-DebtReduction1-1DC-yellowarrow             49        12        0.245
##  2 w-300250-DebtReduction1-1DC-CreditSolutions         77        18        0.234
##  3 w-300250-DebtReduction1-1DC-white                    6         1        0.167
##  4 w-300250-DebtReduction1-1DC                        348        56        0.161
##  5 w-300250-DebtReduction1-1DC-Head2                   89        14        0.157
##  6 w-300250-DebtReduction1-1DC-BlueMeter               92        14        0.152
##  7 w-300250-DebtReduction1-2DC-BlueMeter               87        12        0.138
##  8 w-302252-DebtReduction1-1DC-yellowarrow-bl~        232        31        0.134
##  9 w-302252-DebtReduction1-1DC-CreditSolutions       1054       131        0.124
## 10 w-302252-DebtReduction1-1DC                        272        32        0.118
```

```r
head(campaign_quality, 10)
```

```
## # A tibble: 2 x 4
##   PublisherCampaignName   TotalLeads GoodLeads GoodLeadRate
##   <chr>                        <int>     <dbl>        <dbl>
## 1 DebtReductionCallCenter        271        44        0.162
## 2 DebtReductionInc              2750       349        0.127
```

Observations:

1. **Higher Good Lead Rate for DebtReductionCallCenter**:
   - The `DebtReductionCallCenter` campaign has a Good Lead Rate of approximately 16.24%, which is higher than the `DebtReductionInc` campaign's Good Lead Rate of approximately 12.69%.
2. **Opportunities for Improvement**:
   - Since the `DebtReductionCallCenter` campaign has a higher Good Lead Rate, it suggests that leads from the call center might be of higher quality compared to the leads from the online form submission (`DebtReductionInc`).

Recommendations:

1. **Focus on High-Quality Campaigns**:
   - Given the higher Good Lead Rate from `DebtReductionCallCenter`, consider allocating more resources and efforts towards this campaign to capitalize on its higher lead quality. This could involve increasing the budget for call center activities or enhancing the processes used by call center staff.

2. **Analyze and Improve Online Form Submissions**:
    - For the `DebtReductionInc` campaign, investigate why its Good Lead Rate is lower. Analyze factors such as the form design, the questions asked, the user experience, and the sources driving traffic to the form. Implement improvements based on these findings to boost lead quality.
3. **Enhanced Screening and Follow-Up**:
    - Implement more stringent screening criteria for online form submissions to filter out low-quality leads before they are passed to the advertiser. Additionally, improve follow-up procedures to ensure potential customers are engaged effectively.
4. **Testing and Optimization**:
    - Conduct A/B testing to compare different versions of the online forms and call center scripts to identify the most effective strategies for increasing lead quality.
    - Experiment with different incentives, messaging, and follow-up techniques to see what resonates best with potential leads.
5. **Training and Support**:
    - Provide additional training and support to call center staff to ensure they are equipped with the best techniques for generating and qualifying leads. This could further improve the Good Lead Rate for the `DebtReductionCallCenter` campaign.