REPLACE WITH COVER PAGE

Table of Contents

Introduction

Thank you first and foremost for the opportunity interview with your firm. In the ever increasing discipline of data science it is increasingly more difficult to find the right person with the tools of data science along with the skill of analysis and decision making along the path of the question of interest.

At the onset of this resume being selected to go forward, it is with great respect for the hiring process that the focus remains on the following elements when working through the problem stated:

- Best Practices
- Estimate your model performance on the data provided
- Providing analysis on the thought process along the way

This paper will serve as a summary for the experience with the data provided, the analysis to reach a decision on feature selection, experimentation with various models, and finally choosing a model that aligns with the following business objective (minimizing cost of errors):

- Each False Positive costs $10
- Each False Negative costs $500
- True Positives and True Negatives cost $0

While reviewing technology choices and skills along the way are important to ensure the candidate has a grasp on an approach, this paper serves as a business summary. More details can be found in the GitHub code repository. Generally, there is a high-level framework that most data scientists take when provided a question of interest:

- Step 1: Frame the problem.
- Step 2: Collect the raw data needed
- Step 3: Process the data for analysis
- Step 4: Explore the data
- Step 5: Perform in-depth analysis
- Step 6: Communicate results of the analysis

This paper will take a similar approach above; breaking down each Step into a table of contents.

Data

Naturally, data is probably the most important starting place in the approach to analyze a problem. Data was provided for this task, as such, the steps to view, clean, and analyze the data will be outlined sequentially.

Data Source

For the question of interest, an "UNKOWN" dataset was provided by the hiring manager. Leaving the dataset as "UNKNOWN" ensures no one candidate has a head start in domain expertise. The first step is to approach the data to gain as clear of an understanding as possible given the unknown domain. This usually starts with an EDA process (exploratory data analysis) that can be different based on the features, size, and relevancy of the data.

Exploratory Data Analysis

The EDA task was completed using R Studio. The understanding of the tasks is to classify column "y" using features x0-x49 to minimize costs. Using a skim function; a summary of the data frame can initially be reviewed to determine next steps.

Missing data is one of the first issues a data scientist will likely tackle after gaining access to the data. Initial steps include determining variable means and variable histograms. This dataset appears to be missing data under the triple digit mark which is very few compared to the 160,000 rows of available data. Of note is that the average value of "y" is 0.401, indicating an estimate of 40% of the rows having y=1.

In addition, a review of rows that include "NA" was completed. There is roughly 1,500 rows that are removed that include "NA." This makes up for 0.9% of the data. There certainly are techniques that could be employed to determine imputation standards for this dataset. At this point it is of note to review how these missing data may affect the models once those are explored in detail. In practice, simply taking the median value and applying it for missing values arrives at similar results to the baseline in a model with missing data. This encourages the data scientist to avoid mechanically omitting missing-data records for the sake of saving time and using a few forms of imputation to strengthen the model. At this stage however, the data is simply being reviewed and assessed.

Attributes x37 and x32 are dollar figures and includes a "$" in the cell which is interpreted as a character. This field needs to be converted into a numeric format. A function to index the string with a substring was generated to remove the dollar sign and convert to numeric for both attributes.

```
#col 37 is dollar amounts stored as characters, need to convert
data_v1$x37 <- as.numeric(substr(data_v1$x37, 2, nchar(data_v1$x37))) #index string with substring to cut out the
dollar sign and convert to numeric
#same for col 32
data_v1$x32 <- as.numeric(substr(data_v1$x32, 1, nchar(data_v1$x32)-1)) #same thing
```

In addition to the cleaning of the characters for a numeric field, a function was employed to remove the NA which accounted for another 70 rows that were removed. In this step it was determined that the 70 rows were negligible in the overall dataset. Again, imputing functions could be explored once a model is chosen.

An encoding feature is employed due to a few other variables that require categorical manipulation. The sheer number of categorical variables in each non-numeric column is of interest. Variable x24 appears to be a region or continent and includes "america," "asia," and "europe." Respectively there are

- 4,420 records for "america"
- 137,651 records for "asia"
- 16,381 records for "europe"
- 28 records that were blank.

```
##
##        america    asia  euorpe
##     28    4420  137651   16381
##
##            Apr     Aug     Dev    Feb January    July    Jun    Mar    May
##     30    6701   29125     23    139       9   45140  40922   1221  21711
##    Nov    Oct   sept.
##    332   2385   10742
##
##           friday   monday  thurday  tuesday wednesday
##     30       556      484    29172    27697    100541
```

The same exercise was done for attributes x29 (month) and x30, "day of the week" excluding Saturday and Sunday. Given the domain knowledge of how months, days, and calendar items are categorized, it is important that these remain consistent throughout the data.

```
#day of week encoding
temp_labels <- unique(non_numeric_only$x30)
tc = non_numeric_only$x30 #target column
for(i in 1:length(tc)){
  if(tc[i] == 'monday'){
    encoded_value <- 1}
  if(tc[i] == 'tuesday'){
    encoded_value <- 2}
  if(tc[i] == 'wednesday'){
    encoded_value <- 3}
  if(tc[i] == 'thurday'){
    encoded_value <- 4}
  if(tc[i] == 'friday'){
    encoded_value <- 5}
  if(tc[i] == ''){
    encoded_value <- 6}
  tc[i] <- encoded_value
}
temp_labels
```

```
## [1] "tuesday"   "wednesday" "thurday"   "monday"    "friday"    ""
```

Numeric values were assigned to each region:

- "america" = 1
- "asia" = 2
- "europe" = 3
- blank = 4.

A "region_numeric" data frame is created for later use in the model.

Monthly encoding is also performed to convert variable x29 (months) into numeric form. A "month_numeric" data frame is created for later use in the model building. Finally the same process is completed for attribute x30 (day of week). Similarly, a "day of the week_numeric" data frame is created to use in the model building step. Each of these categorical changes were vetted by backing into the data using a table function to produce the output before and after the conversion for repeatability in the data (if new data was provided later).
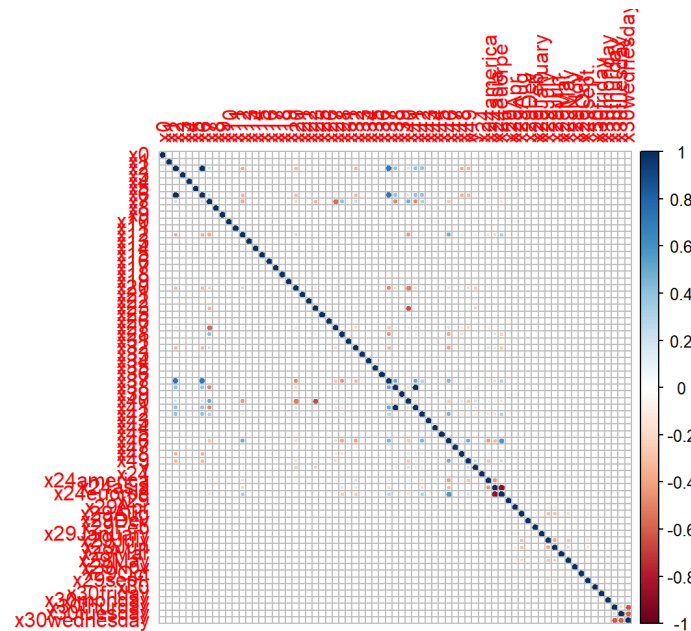
Dummy variables were then created for each of the categorical variables. This required four new variables for region, 13 new variables for month, and 5 new variables for day of the week. There are 23 regional values and 12 days of the week values that are blank.

Correlation Analysis

After cleaning the data, the next step is to evaluate correlation between the data to narrow down a model approach for prediction. Attributes x24, x29 and x30 are region, month and day of the week (weekday only) respectively.

Each of these went through one-hot encoding which generated new dummy variables for binary values in each category. About 0.9% of the is data missing with minimal value from imputing; therefore these rows values were removed from the data set. The values were imputed into the new data frame including the new dummy variables in order to generate correlation analysis.

According to the correlation plot of all the variables, there are appears to be strong negative correlations between Wednesday and Tuesday, as well as Asia and Europe. There does not appear to be strong correlation among any of the individual categorical variables and target y.

Reducing the correlation matrix down to the top 20 variables provides a much more clear view of the relationship.

Model 1

Lorum ipsum dolar

Results

Lorum ipsum dolar

Summary

Lorum ipsum dolar

Model 2

Lorum ipsum dolar

Results

Lorum ipsum dolar

Summary

Lorum ipsum dolar

Model 3

Lorum ipsum dolar

Results

Lorum ipsum dolar

Summary

Lorum ipsum dolar

Model 4

Lorum ipsum dolar

Results

Lorum ipsum dolar

Summary

Lorum ipsum dolar

Conclusion

note, store expected costs from model classifications for graphical comparison at the end??
viz ideas - fancy colored table for confusion matrix/price matrix stacked bar chart for each
model, number of type1/type2 errors and the total cost for 3 bars on each model

Recommendation

Lorum ipsum dolar

Appendix

As with any technology; the system required to generate meaningful data, cleaning of the
data, as well as the models that were implemented require various software packages. The
raw "code" used to explore the methods detailed above can be found in the following
libraries:

- GitHub Repository
- R Markdown Results
- R Markdown for Random Forest Only