## Red Ventures Data Science Technical Interview Study Case

## Introduction:

We would love to understand your data and machine learning technical skills, critical thinking on the business problem and insights at hand, and ability to derive actionable insights.

Through this case, please provide a notebook of your choice (Jupyter, Colab, etc.), including your code, your written analysis and explanation for each question, and any visualizations you deem appropriate or relevant. Use appropriate modeling techniques, data analysis, and visualization techniques to support your findings. There is an appendix with images and a data dictionary at the end of the doc.

**You are strongly encouraged to submit different iterations of your analysis and modeling. This will help the code reviewer understand your complete thought process, including what worked, what didn’t, and the decisions you made throughout, as the evaluation is not solely based on the model’s performance. Ensure your submission is formatted in a way that is easy to follow.**

Thank you in advance for participating in the case study, and enjoy!

## 

## Timing:

For your time, we recommend you do not exceed 4-5 hours on the case.

## Case Study:

### Description

Financial Services stands out as the premier financial authority. When you visit Financialservices.com, the reviews, guides, and educational content have been developed by leading personal finance experts. Financial Services’s product comparison tools, calculators, and educational content help over 100 million consumers make smarter financial decisions each year. No matter where you are in your financial journey, Financial Services can help you reach your goals.

One particular area in which Financial Services places a strong emphasis is mortgages, with the goal of advancing the visitor's decision-making journey and ultimately guiding them toward applying for and securing a mortgage. The mortgage team’s goal is to **maximize revenue** generated from users on our website by getting them to **schedule an appointment** with one of Financial Services’ mortgage partners. Once the appointment is scheduled, Financial Services gets paid a bounty (revenue). The amount depends on the variation of mortgage chosen for the appointment. For the purpose of this case, there are 4 variations (A, B, C, D), each with their own bounty associated with them.

The customer journey is straightforward. A user visits a page on the website. Once on the page, they have the option to click on a banner that takes them to a form that they can fill out to schedule an appointment with a mortgage lender. They have the option to choose from four mortgage types when they schedule the appointment.

To support this goal, the mortgage team conducted a split test, also called A/B testing, on various mortgage pages. This testing allows the team to compare the performance of website creative variations to see which one appeals more to visitors to maximize a targeted metric. The objective of this test was to learn which banner’s **call-to-action (CTA)** title copy and on-page placement combination will best entice visitors to click and enter the scheduling form in hope of increasing appointments and overall revenue. *See appendix for visual samples of each CTA and placement variation.*

### Data

The testing has concluded and we have a data set of 120,000 decisions (rows) of the combination of CTA Copy and CTA placement being served randomly to users coming to the website. To help increase revenue per decision, the team believes that certain groups of people may have performed differently based on the CTA copy and placement. We can use an algorithm to predict in real-time what CTA combination would be best to show to a future user to maximize the key metric.

For assessment purposes, these data have been partitioned into 2 sets:

* **train.csv**
* **test.csv**

Train.csv includes 100,000 rows of the original data with labels, test.csv contains 20,000 rows of the original data without labels. You will use test.csv to make predictions for submission.

## 

## Tasks

### Part 1 Predictive Model Building

You are asked to build a predictive model to make predictions on an unlabeled test set test.csv. Please review the dataset carefully to select the appropriate columns for your model building.

The predictions should be the probability of **ClickedCTA** Pr(ClickedCTA = 1) in [0, 1].

**Evaluation Metric**

Log Loss will be used to score the fitness of your classifier. In short, log loss penalizes models that are highly confident about wrong answers. 0 is a perfect score. The train() function from caret in R, and sklearn.metrics from python both include a log loss metric for model optimization.

### Part 2 Answering Questions

Below are a few questions that would help you gain a better understanding of the dataset and the business context:

1. *What relevant key metrics are provided to evaluate the CTA combinations? And which CTA Copy and CTA Placement did best/worst based on the key metrics?*
2. *Which groups of people tend to be more correlated or less correlated with our key metrics?*
3. *What ways can you manipulate the columns/dataset to create features that increase predictive power towards our key metric?*
4. *Besides Log Loss, what other metrics will you use to evaluate the model's performance, and why?*

#### Optional:

If you have more time, please answer the following questions:

1. *What additional predictive model would you build to inform which CTA combinations would maximize the revenue in addition to predict ClickedCTA, and why?*
2. *If we called one of these CTA combinations our champion (serve it 100% of the time), how much incrementally is that worth to us vs. the average of the rest of the split test?*

Please provide your responses to the above questions as the last section of your code script.

Submitting Predictions

As part of your submission archive, create a CSV with two columns **[userId, pr\_CTA]** that contain your predicted probabilities for each userId in the test set. The file-name should follow the format **<name>\_predictions.csv**, for example, **LastName\_predictions.csv**. Three files are included to help ensure your predictions are in the correct format:

* example\_predictions.csv
* make\_submission.R
* make\_submission.py

example\_predicions.csv is an example of what your submission CSV should look like. Additionally, convenience functions for both Python and R have been provided to turn data frames into appropriately formatted CSV.

## 

## Deliverables

You should submit your deliverables as a single, compressed archive, labeled <name>.zip. For example, **FirstName\_LastName.zip**.

You will be assessed in two ways. First, you will be objectively scored on how well your classifier is able to predict a user’s likelihood to click on CTA. Secondly, you will be assessed on the quality of your methodology and ability to describe and justify your design decisions and questions answering.

Your submission archive should contain:

1. <name>\_predictions.csv Your predictions on the test set. Format is described below.
2. code/ a directory containing the code you used to make predictions. This code should be reproducible. That is, given your code and the train and test data sets, we should be able to exactly reproduce your predictions CSV.

This archive should be emailed to the recruiter or uploaded to the shared Google Drive folder if you have the link to that folder.

## APPENDIX

### CTA Versions

| Version 1 |
| --- |
|  |
| Version 2 |
|  |
| Version 3 |
|  |

### 

### Page Placements

| Placement 1 - Top | Placement 2 - Middle | Placement 3 - Bottom |
| --- | --- | --- |
|  |  |  |

### Data Dictionary

| **Variables** | **Definition** |
| --- | --- |
| userId | Unique identifier for users visiting Financial Services |
| sessionReferrer | Source from which the user arrived (e.g., search engine) |
| browser | Browser used by the user (e.g., Chrome, Firefox) |
| deviceType | Category of device used by the user (e.g., mobile, desktop, tablet) |
| estimatedAnnualIncome | Estimated annual income of the user based on geographic location |
| estimatedPropertyType | Estimated property type (e.g., residential, commercial) |
| visitCount | Number of previous visits did the user have before the current visit |
| pageURL | URL of the current page visited |
| ctaCopy | Text of the call to action that was prompted to the user for potential engagement |
| ctaPlacement | Location of the call-to-action button on the page |
| editorialSnippet | Text surrounding the call-to-action |
| scrollDepth | Percentage of the page length scrolled by the user.  If the user didn’t scroll on the page AFTER the ctaCopy and ctaPlacement already loaded, than scrollDepth will be 0. |
| clickedCTA | If the user clicked the call-to-action (yes 1/no 0) |
| submittedForm | If a user submitted the mortgage application form (yes 1/no 0) |
| scheduledAppointment | If the user scheduled an appointment regarding the mortgage application (yes 1/no 0) |
| mortgageVariation | Mortgage product variation the user applied for to discuss during the appointment. This only is present when someone scheduled an appointment |
| revenue | Revenue Financial Services received from the appointment scheduling and mortgage variation. If no scheduled appointment, this will be 0. |