## Scenario



You are part of a research and modeling team at National City Bank. You team has been asked to create a customer propensity model for a new product, specifically a line of credit against a household’s *used* car. Since the line of credit product is only in pilot, you are asked to identify the next 100 customers from a prospective customer list to contact. Bankers will call and direct mail will be sent to households your model identifies with the greatest probability of accepting the offer. Once your team has modeled and identified the customers, you must present your findings to the bank’s chief product officer. Once she/he feels comfortable with your proposal, marketing will begin its process.

**You are asked to examine the historical data from 4000 previous calls and mailings for the line of credit offer. Using this historical data, and any supplemental data, perform EDA, create a propensity model, evaluate it and identify by uniqueID the top 100 households to contact from the prospective customer list. Additionally, bank executives are eager to learn more about the customer profile for historical and top prospective customers. As a result, variable importance and sound EDA will aid the presentation. You will need to turn in code and PowerPoint slides.**

## Data

*This is the original data source but is not the data to be used in the case. The case data is stored in the repository.*

Source: <https://www.kaggle.com/kondla/carinsurance>

Supplemental data represents fictitious 3rd party data that the bank would purchase to improve the model’s accuracy.

## Example *Abridged* Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| HHuniqueID | Communication | LastContactDay | LastContactMonth | CallStart | … | Y\_AccetpedOffer |
| HHd4d0af8c72 | telephone | 28 | jan | 13:45:20 | … | 0 |
| HH8d3e87c164 | NA | 26 | may | 14:49:03 | … | 0 |
| HHdd53ef1db6 | cellular | 3 | jun | 16:30:24 | … | 1 |
| HH6fa0de6516 | cellular | 11 | may | 12:06:43 | … | 1 |
| HHeb436ca7cf | cellular | 3 | jun | 14:35:44 | … | 0 |
| HH5119beb3cd | cellular | 22 | may | 14:58:08 | … | 1 |

## The Submission

* The submission will include business analyst slides covering the problem, data, methods, model explanation and any insights. Without a presentation, the “organization” section of the rubric will be 0. Exceptional submissions are well ordered and provide a coherent narrative covering data, methods, modeling and any insights that may be of interest to the audience.
* The submission will include a supplemental identifying the top 100 households by ID and corresponding probabilities. This can be in a CSV or similar file format. In addition, any insights identified in the presentation will be included in a written supplemental. The insights written portion can be 3-5 sentences for each insight in a bulleted list format. Exceptional submissions include statistics from external credible sources that support the identified personas or insights. For example, “…fewer calls to landlines are successful in the month of XX, one explanation may be that the Bureau of Labor Statistics shows that people are more likely to vacation in this month…”. Without a the top 100 households AND a written supplemental that coincides with the narration and supported by code the “written supplemental” section will be 0.
* The submission will include either a recorded screen narration of the business presentation, a text file with a URL to a recording (like youtube video) or audio that is embedded into the slide deck. Tone, volume, cadence, use of filler words and pronunciation will be accounted for in this section. No points will be deducted based on English proficiency (ie ESL) but technical descriptions that are incorrect will be detrimental such as “Logistic Regression is used for predicting continuous outcomes”. Failure to submit a narration, the “delivery” section of the rubric will be 0.
* An R script covering all data munging, modeling, evaluation and visualization construction used to create the presentation artifacts (you do not need to use R to construct the slides but it is possible) and come to the case outcomes. Your code must use the following R functions at least once throughout your code, group\_by, aggregate & subset in addition to modeling code. Make sure to that your code contains ample comments. Failure to turn in an R script will result in a “Documentation” score of 0.

## Criteria for Success

The presentation will be evaluated on an equal weighted scale with the following criteria. For example 20 points per each category [depends on the individual course weighting found in Canvas]

* **Organization** – Was the presentation well organized?
* **Delivery** – Was the content delivered clearly and persuasively with the audience in mind?
* **Code Documentation** – Was the data mined to support the conclusion?
* **Written Supplemental** – Are the bullets clear and supported in narration and code? Were the top 100 households identified?
* **Data Mining & Modeling Proces**s – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively? Was CRISP-DM, SEMMA or similar followed to organize the work?

## Another resource may be a public kaggle kernel

*Keep in mind this may not be helpful but code can be examined for additional ideas.*

<https://www.kaggle.com/kondla/simple-random-forest-on-insurance-call-forecast/code>