GRA 41363 Machine Learning for Business - Exam description

You work as a data scientist for a large US-based real estate company, *Rockafella Properties*. The firms' competitors have started offering home insurance to owners. In response, your firm has decided to offer home insurance to new owner-occupied properties, under a new division called *Rockafella Insurance*. Home buyers would be free to decide to purchase home insurance or not. Your boss has informed you that you can use historical data surveyed and documented by the 2011 National American Housing Survey (AHS) for this task. He has hinted that you can use 2011 AHS data on owner- occupied housing characteristics, housing costs, household demographics, and whether the household has homeowner insurance (the BUYI variable) to predict (using an ML model) who would be a good target for Rockafella's agent to prepare an insurance offer. He has provided you with a sample of the data. This sample is available in the google drive folder below in a file called"ahs_insurance_sample.xlsx". For logistical reasons, your company can consider offering home insurance to 5% of new owner-occupied home buyers (i.e., who home buyers who intend to live in their new home). Every year in the US, 4 million new homes are sold to owner-occupied home buyers.

The data sample is provided here:

https://drive.google.com/drive/folders/164rzORHTegJ5XAAvC2w7baWQ2WEDdPpf?usp=sharing

A mini codebook has been shared with you in the data folder which provides variable descriptions for the variables you have been provided (and extra variables that you can ignore). You can find the definition of values that the features take (as "response codes") and the variable descriptions on this webpage, and can perform a Keyword Search (note the "Search For" field):
https://www.census.gov/data-tools/demo/codebook/ahs/ahsdict.html?s_year=2011%20National

(**Note:** This is the data dictionary for the 2011 National AHS. In the data, note that the following codes are recorded for the survey responses: -9 or Blank = Not reported; -8 = Refused; -7 = Don't know; -6 = Not applicable; From the AHS data, only the BUYI variable has been recoded from its original coding (seen in the census.gov link) to 0 if an individual didn't buy home insurance, and 1 if they did.

Your job is as follows. Use the provided data and data dictionary to build a machine learning model predicting whether a new home buyer will buy home insurance from Rockafella Insurance, and estimate the **annual profitability** of deploying your ML model.

For purposes of the exam, assume:

a) When Rockafella Insurance correctly predicts someone to buy home insurance, it earns a 30% profit margin, on the Annual Cost of Homeowners Insurance to the home buyer (recorded as AMTI, in the dataset)

b) It takes an hour for a Rockafella agent to prepare an insurance offer

c) This preparation only results in labor costs for the agent, equal to 500USD/hour.

d) Failing to prepare an offer for a home buyer who would have bought it, will cost Rockafella
by reducing its market share and pricing power over the long run. The cost of this is assumed
to be 200USD for each such failure.

e) Currently your firm does not offer any home insurance. Because all competitors are starting
to offer home insurance to the same 5% of population you can target, by not offering home insurance, Rockafella is currently incurring the same loss to its market share and pricing power, that is mentioned above. So, see item d) above for estimation of this cost per home buyer.

f) Beside those indicated above, there are no costs, for predicting home insurance buyers. There are no benefits, beyond those mentioned above, for Rockafella home insurance or properties businesses.

If you need to make other assumptions, feel tree to do so. Just note the assumption and any references you use to ground the assumption (if necessary).

DELIVERABLE

Write a report based on your findings (5 pages max **and** 1800 words max; “**and”** means both limits apply; This includes front page, table of contents, reference, and figure list as well as any appendices, while no appendix is necessary). This should include:

1. A concrete and valid **recommendation** (specifying the baseline, an appropriate action or actions and well-motivated prediction thresholds) and the associated **total profit**
2. The **action/intervention** on which you base your profit calculation
3. Your **data management procedures**, consisting of:
    - Your **review of the data**;
    - Any **preprocessing** and **feature engineering** (Do these in Excel, Python or other
      appropriate programs) steps; Note that a basic level of preprocessing and feature engineering would have to be done. Motivating the rational and implications of consequential decisions you make here are required and valuable; e.g. these could be motivated by summary statistics and data explorations.

i. To be specific a basic level of preprocessing and feature engineering, would constitute some of the following (among others referred to in class), where appropriate:

1. Feature creation
2. Feature exclusion (redundant features, etc.)
3. Encoding (if appropriate)
4. Row exclusions (if appropriate)
5. Outlier removal (if appropriate)
6. Numerical transformations (if appropriate)
7. Dealing with excess zeros, missing values, etc.

c. Your **employed data partitioning** and **justification** of it;

- The composition of the **profit components** (e.g., what are your high-level profit components, and how does each contribute to the total?). Ensure to specify qualitative and numerical representation of each component of a profit matrix (the numerical representation could be a formula or a number; you are to decide how to represent it.)
- **Technical/data issues** you think might affect results

4. Your **modeling procedures**:

a. Your **model selection process**, that:

- **i.** **Searches across possible models and hyperparameters (this can be done using automated tools, such as DataRobot or tpot)**
- **ii.** **Searches across meaningful sampling routines (e.g. downsampling or SMOTE)**
- **iii.** **Searches across potential prediction decision thresholds**
- **iv.** **Searches across potential actions/targets**

5. Your **evaluation process**, that:

- Executes the above without compromising the holdout, overfitting or failing to address any leakage
- **Describes which features hold the most signal and motivate reasoning**
- **Makes correct use of, and interpretation of, partitioning decisions**
- Addresses other key modelling issues

6. **CRUCIAL NOTE:** You **must begin all paragraphs with a 1- to 7-word title that describes the paragraph** (the purpose of this is to provide more structure, cohesion & flow to your writing), an example would be: "***On data splitting:*** *We split the dataset into three subsets: training, valuation and holdout. We used 60% of the data for training, 20% for valuation and 20% for testing."*
7. Clear and organized reporting of machine learning model prediction and evaluation steps

TIPS, TRICKS, and NOTES

1. I strongly suggest using the same structure, as the enumerated list above. That means use a larger font heading for sections corresponding to **words in bold font** in the “deliverable” instructions above belonging to numbers 1 to 6, and smaller headings for subsections corresponding to words in bold belonging to a) to e), and yet smaller headings for the subsubsections, i) to iv).
2. The **words in bold font** in the “deliverable” instructions above, are great (not exhaustive) suggestions for the titles (or part of titles), which each paragraph in the deliverables is to start with.
3. You may use Python (in conjunction with Excel), DataRobot (in conjunction with Excel), or some combination for the exam. If using DataRobot, choosing “Quick” as the

modeling mode (under
Start), will be adequate.

4. You will not be graded based on your solution's profit. If one student's model yields expected profit
X, and another student's model yields expected profit 10*X, the 10*X exam will not necessarily receive a higher score. Exams will be graded on their model building, model evaluation, and model valuation processes. Solution profitability may spuriously correlate with exam grades to the extent that a more thorough modeling procedure may yield a more profitable solution.
5. The data contain ~50K rows. Be aware of your time limit, taking into consideration the modeling platform (Python, DataRobot) and models that you build, and your resource limitations (hardware, memory, etc).
6. You are not to submit Python codes (references to code are fine). However, where Python is used for modelling, explaining the modelling procedure(s) is a requirement.
7. I strongly recommend that you do not rely on online solutions for similar code. Using existing code runs a risk of yielding a good solution without demonstrating that you know how to properly execute the modeling process. Because you will not be graded on your model's profitability (the goal of many online codes), but will be graded based on the thoroughness of your modeling process and explanation of model value, adapting existing solutions online can often become a minefield of shortcuts and misunderstandings that hurt exams.
8. Many python machine learning models and methods may require your data to be encoded (e.g. one-hot encoded instead of categorical values). In that case, either **implement the correct encoding or ensure to find and apply the equivalent method for your data type. Ensure to do so in DataRobot too.** You can find the appropriate methods online (by googling or searching in documentations for packages we have introduced you to).
9. If you modify the dataset in Excel, it is possible that automated formatting changes could make the dataset un-loadable into DataRobot. This is usually a function of semi-colons, commas, and decimals being used differently around the world. If you find that a dataset you've modified in Excel cannot be loaded to DataRobot, you may need to come up with a workaround. If this is a

problem for you, I often find it easiest to load Excel data into Python, save data from Python into .txt or .csv, and then load that into DataRobot.

10. A recommendation given this exam has a time limit: In exams, work projects, or your thesis project,
    always get 1 “OK” solution first. Even if you immediately envision a grand solution and the “OK” solution seems like a wasted interim step, start with a simple but informative approach! You’ll learn something from this process, and you’ll ensure that you’ve got something to hand in by the deadline.
11. You will be graded on the first 1800 words (and 5 pages) of your report. Title words do count towards the total word count. The limits are meant to help replicate reports to technically-oriented business managers and data scientists in firms.