

**** Candidate Instructions

****

Case Study – Claims Fraud Detection

Contents

[INTRODUCTION 3](#_gjdgxs)

[PROJECT DESCRIPTION 5](#_30j0zll)

[DATA 6](#_3znysh7)

[PROJECT REPORT 1](#_tyjcwt)1

INTRODUCTION

Congratulations, you are on the final step to earning your CSPA designation! All other CSPA requirements must be completed (or waived) before you begin this case study project.

Overview

The objectives of the Case Study are:

* Demonstrate your ability to apply the knowledge and skills covered in the other CSPA requirements
* Build a predictive model to design a solution for the given project scope, business problem, and multiple datasets
* Document your modeling process, in a project report including approach, programming code, and data visualization
* Complete the project within a 60-day time period
* Submit your report and scored output to be reviewed by the iCAS
* Respond to any questions about your project, if requested, in order to complete the project requirements

You have agreed that this project will be done by you independently with no hands-on assistance. The use of online resources, such as but not limited to Google, Stack Overflow, Reddit, r-project, and Python.org, is permissible. You may also consult with colleagues for general questions and advice; however, these instructions and the data sets are proprietary to The CAS Institute and should not be shared. While Excel can be used to perform rudimentary tasks involving the data, we expect you to use Python, R, or SAS to do exploratory data analysis, data wrangling, dataset merging, model development, and model performance diagnostics. All code, documentation, exhibits, and the final report must be written by you. Your submitted code should be able to reproduce your data scrubbing and modeling results programmatically. If, during the course of the project, you feel that you are in need of mentorship, please contact iCAS Director Alicia Burke at aburke@casact.org. You have been provided with the necessary datasets (explained in more detail below), a Final Report template, a Scored Excel workbook, and these instructions.

The project report is to be prepared using the ‘Final Report’ template provided to you. The report is limited to 3,000 words, not including any exhibits, code, charts, or tables you deem necessary to support your work.

Project Submission

The assessment results are to be entered into the Microsoft Excel file (.xlsx extension) named ‘Scored’ that has three columns – Claim\_ID, binary prediction, and score underlying the prediction (raw model output).

When finished with model building, please save your Final Report (do not rename it), your Scored.xlsx file (do not rename it), and all of your final code must be in a single file named final\_code.R, final\_code.py, or final\_code.sas (Ex: If you wrote 3 codes; one to do EDA, one to develop the model, and one to do model diagnostic, we want a single code file that contains all 3 codes used in the right order of execution) and submit to The CAS Institute. These files must be submitted or there will be automatic failure of the project. Please also submit any other exhibits or documentation you want considered in grading your project. The grader will have access to all files you submit and may review them as necessary. Please do not put your name on any of the files you submit and ensure that any and all personally identifiable information is removed. This includes removing your name or employer name from file paths and removing your name from any metadata in Excel and Word documents. Information on how to remove personally identifiable information from Excel and Word documents can be found at this [link](https://support.microsoft.com/en-au/office/help-protect-your-privacy-252a47ec-1b31-4fd0-8450-e66d6c2de950#:~:text=If%20you%20share%20a%20document,the%20Word%20menu%2C%20click%20Preferences.&text=Under%20Privacy%20options%2C%20select%20the,file%20on%20save%20check%20box.). When you are ready to submit your Final Report and Scored.xlsx files, as well as any additional files you wish to submit, for grading, upload your files to the Output folder in the same location as the Input folder from which you downloaded the assigned datasets and other files, including these Instructions.

PROJECT DESCRIPTION

For this project, you are a data scientist in Company ABC that writes auto insurance. The claims department has reached out for guidance. Detecting fraudulent claims has always been tricky and with the onset of predictive modeling, they are hoping you can build a model to combat the situation. In this Case Study, you are to build a predictive model (or several if needed) to detect whether a claim is fraudulent or not. The dataset provided contains characteristics of historical claims, policy characteristics, driver characteristics, vehicle characteristics, and most importantly, **the claim’s investigation results on whether the claim was fraudulent or not (which would serve as the target variable of your model).**

**Since investigations are costly, the stakeholders require that at least one claim be fraudulent for every four that is flagged by the model. That is, due to resource constraints, the Company wants to be reasonably sure that the model they are implementing is confident in labeling a claim as fraudulent.**

When finished with model building, please follow the instructions outlined in section “Project Submission”.

DATA

The data used in this case study was simulated and should not be relied upon in any way for any use outside of this Case Study project. The distributions, relationships, correlations, and lack of relationships in the data are not from actual data. The sole purpose of this simulated data is for this CSPA Case study.

You are given five datasets, four of which are for the modeling process and one for scoring:

* “Claims” contains data related to the specific Claim\_ID, including “Fraud\_Ind” which states whether the claim was fraudulent or not
* “Driver” contains data related to the driver
* “Policy” contains data related to the policy, including historical claim experience
* “Vehicle” contains data related to the vehicle
* “Assessment” contains data from the above-mentioned datasets, but without “Fraud\_Ind”

It should be noted that you may not be provided with all variables listed below in the data dictionary. Furthermore, the actual data and what is described below may be different. You are to build the best model you can with the data provided to you.

Claims

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Variable Type | Variable Range | Variable Category |
| Claim\_ID | Unique identifier for claims | ID | N/A - Not a variable | Claim |
| ClaimAmount | The dollar amount associated with the claim ID | Numeric | At least 0 | Claim |
| Claim Notes | Notes from the claim adjuster | Categorical | Free Text | Claim |
| Report\_Lag | How long it took for insured to report the claim (in days) | Numeric | At least 0 | Claim |
| Fraud\_Ind | Whether the claim is flagged as fraudulent | Categorical | 0; 1 | Claim |
| Accident\_date | Date of the accident | Date | Date | Claim |
| Report\_date | When the insured reported the accident | Date | Date | Claim |

Driver

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Variable Type | Variable Range | Variable Category |
| Driver\_ID | Unique identifier for drivers | ID | N/A - Not a variable | Driver |
| Education | Highest level of education of insured | Categorical | High School or GED; Bachelors; Masters; Some High School; Doctorate; | Driver |
| Yrs\_Licensed | Years licensed of driver | Numeric | 0 to 88 | Driver |
| Marital\_Status | Marital status of the insured | Categorical | Single; Married | Driver |
| INCOME | Income level of insured; Poverty = below poverty line, Working = low income, Middle = middle class, Upper = upper class | Categorical | Middle; Working; Poverty; Upper | Driver |
| CREDIT\_SCORE | Credit score of insured, ranges from 0 (bad) to 1 (good) | Numeric | 0 to 1 | Driver |
| Cell\_Usage | Cell phone usage in the vehicle by insured | Numeric | At least 0 | Driver |

Policy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Variable Type | Variable Range | Variable Category |
| Altitude | Altitude at insured’s garage location (in meters) | Numeric | At least 0 | Policy |
| Foggy\_Days | Average number of foggy days in the past year at insured’s garage location | Numeric | At least 0 | Policy |
| Clms\_flt# | Number of historical at-fault claims in the past # year(s) for the insured | Numeric | At least 0 | Policy |
| Clms\_naf# | Number of historical not at-fault claims in the past # year(s) for the insured | Numeric | At least 0 | Policy |
| Viol\_mjr# | Number of historical major violations in past # year(s) by insured | Numeric | At least 0 | Policy |
| Viol\_mnr# | Number of historical minor violations in the past # year(s) by insured | Numeric | At least 0 | Policy |
| Late\_90d | Number of late payments in the past 90 days on insured’s credit report | Numeric | At least 0 | Policy |
| Num\_Accts | Number of open accounts on insured’s credit report | Numeric | At least 0 | Policy |
| Outs\_Bal | Amount of debt on insured’s credit report | Numeric | At least 0 | Policy |
| Pop\_Density | Population density of area around insured’s garage location; 1 is low, 4 is high | Categorical | 1; 2; 3; 4 | Policy |
| Seat\_Belt | Seat belt usage per week of the insured in the vehicle | Categorical | Usually; Occasionally; Always; Rarely; Unknown; Never | Policy |
| Time\_bet10pm2am | Average percent of time insured spends driving between 10 p.m. and 2 a.m. | Numeric | 0 to 1 | Policy |
| Time\_highway | Average percent of time insured spends on the highway | Numeric | 0 to 1 | Policy |
| Exposure | Number of car years (i.e. the portion of a full year the policy is in force. 1 if the policy is in force for a full year.) | Numeric | 0 to 1 | Policy |
| Policy\_ID | Unique identifier for policies | ID | N/A - Not a variable | Policy |
| Policy\_Orig\_Eff\_date | Policy Effective Date (Original inception, may not be current term) | Date | Date | Policy |

Vehicle

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Variable Type | Variable Range | Variable Category |
| Braking\_Mile | Average number of hard braking events per mile by insured | Numeric | At least 0 | Vehicle |
| HP\_Vehicle | Horsepower of the vehicle the insured drives | Numeric | 40 to 980 | Vehicle |
| Make | Make of the vehicle the insured drives | Categorical | Many | Vehicle |
| Model | Model of the insured’s vehicle | Categorical | Many | Vehicle |
| Model\_Year | Model year of the insured’s vehicle | Numeric | 1923 to 2024 | Vehicle |
| Left\_Mile | Average number of hard left turns per mile by | Numeric | At least 0 | Vehicle |
| Vehicle\_ID | Unique identifier for vehicles | ID | N/A - Not a variable | Vehicle |

Assessment

This dataset is not to be used in model development or model validation – view this data as an implementation of your model in practice at your Company. It contains the same information as Claim, Driver, Policy, and Vehicle, but on a different group of claims. Upon completion of the model build and validation process, you are required to use your model and provide prediction (i.e., scores) for each claim in the Assessment data. The Assessment results are to be entered into pre-loaded Microsoft EXCEL file (.XLSX extension) named ‘Scored’ that has three columns – Claim\_ID, binary prediction, and score underlying the prediction (raw model output). The Reviewers will use the ‘Scored’ data to evaluate your model.

PROJECT REPORT

In your final report, please reference the ‘Final Report’ template found in your assigned Input folder. In your final report, you should provide background on data preparation and exploratory data analysis, model development, results, and conclusions with sufficient clarity that the reviewer can make an objective appraisal of the reasonableness of your work. The documentation should be sufficient to demonstrate a candidate’s ability to apply the knowledge and skills covered in the other CSPA requirements.

After a committee has evaluated your report, you will be notified if you successfully passed, if additional work is needed to complete the project, or if your project was not successful at meeting the minimum requirements needed to receive a second attempt. Moreover, if we notify that additional work is needed, note that second submissions are final.

Candidates will be evaluated in the following areas, with a minimum proficiency in each area being required to receive a passing grade.

Exploratory Data Analysis and Data Preparation

* Correctly merged and joined datasets as necessary
* Demonstrated an understanding of issues with different data types
* Determined appropriate target and predictor variables
* Provided programming code used to conduct exploratory data analysis
* Demonstrated reasonable exploratory analysis and ability to query the data
* Cleaned miscodings and unobservable values
* Detected outliers and data errors
* Handled high cardinality variables through use of splines or bins
* Addressed missing values
* Assessed and appropriately handled linear and non-linear correlations
* Appropriately separated data into model build and hold out samples (training and validation data)

Model Development

* Chose a reasonable approach for modeling
* Applied proper use of analytical techniques
* Thoroughness and breadth of techniques used
* Provided programming code used to build the model
* Refined approach by performing numerous model iterations and comparing results
* Demonstrated reasonable variable selection
* Model parameters and coefficients made sense for the business problem

Model Performance

* Chose a reasonable approach to validate and test the model
* If alternative models or versions were built, performed appropriate comparisons
* Used reasonable metrics to measure model performance
* Provided programming code used to evaluate model performance
* Provided reasonable and appropriate types of visualizations
* Model performed reasonably well on the hold out data set

Final Report

* Formatting follows requirements:
  + Maximum word count of 3,000
  + Language must be English
  + Outside the word limit can be any number of charts, tables, graphs, dashboards, diagrams, flow charts, programming code, etc.
* Presentation was coherent and organized in a logical and professional fashion
* Provided a clear understanding of the business problem
* Provided reasonable rationale for the selected modeling approach
* Documented material assumptions in sufficient detail to assess reasonableness
* Disclosed potential flaws of the model approach
* Provided reasonable interpretations of results