Всем доброе утро!
Домашнее задание №3

Необходимо построить классификационную модель, оценивающую вероятность дефолта клиента на стадии заведения кредитной заявки.

Для этого необходимо:
0. Скачать данные по ссылке https://drive.google.com/drive/u/0/folders/16CMyPnLu7Fv7IgsYOZimQK-7MaFZEWEZ
Каждый студент выбирает 2 выборки "accept" и "reject", которые начинаются с варианта студента.
Номера вариантов для ДЗ №3 необходимо взять такие же как и для ДЗ №1

Выполненное задание строго необходимо отправить в следующем виде:

1) Файл/скрипты с построенными моделями (обязательно должны быть комментарии, без комментариев задание считается нерешенным)
2) Excel файл с ответами на следующие вопросы:

1. Какая доля 1 в выборке "accept"?
2. Необходимо рассчитать для всех интервальных переменных следующее:
  - Доля пропущенных значений
  - Медиана
  - Среднее
  - Среднеквадратическое отклонение
  - Information Value для каждой переменной
3. Необходимо рассчитать для всех категориальных переменных следующее:
  - Мода
  - Доля пропущенных значений
  - Information Value
  - Есть ли выбросы, аномальные значений
4. Построить логистическую регрессию только на одобренных заявках с преобразованными переменными WoE. Какое значение GINI? F1 мера?
5. Провести анализ Reject Inference. Какая доля отказанных заявок от всей выборки заявок?
6. Построить логистическую регрессию на всех заявках с преобразованными переменными WoE. Какое значение GINI, F1? Изменилась ли модель?
7. Какую модель вы рекомендуете для внедрения в продуктивную среду? Дать развернутое пояснение

Оценка за домашние задания №3 выставляется по 2-балльной шкале, где «2» — задание решено полностью, «1» — задание решено не полностью или с недочётами, «0» — задание не решено или решено неверно.
Перевод оценки за домашние задания из 2-балльной шкалы в 10-балльную проводится путём умножения оценки на 5 без округления.

За домашнюю работу №3 будут выставляться оценки:
«2» – правильно построена модель и на выборке accept, и на выборке reject. Даны верные ответы.
«1,6» – правильно построена модель и на выборке accept, и на выборке reject. 50% ответов верные.
«1,4» – правильно построена модель и на выборке accept, и на выборке reject. Даны неверные ответы
«1» - построена модель только на выборке accept.
«0,8» - задание решено не полностью. 50% ответов верные.
«0» — задание не решено или решено неверно.

Срок сдачи – до 28.03.2022 00-00 (Московское время) ключительно

Решения присылать на почту Maria.Vorobyova.Ser@gmail.com и msvorobeva@hse.ru
В теме письма обязательно должно быть следующее: ВШЭ + Номер курса+номер_вариант+ФИО.

В названии файла необходимо указать:
*Номер курса
*Номер варианта
*ФИО

Пример: «ПМИ_3курс_Вариант_8_ИвановИванИванович»

Если работы будут повторять друг друга, обе работы будут считаться нерешенными.

Homework #3

You should develop a classification model that estimates the probability of a client’s default on the stage of a loan application.

Next steps:
0. Download data from the link https://drive.google.com/drive/u/0/folders/16CMyPnLu7Fv7IgsYOZimQK-7MaFZEWEZ
1. Each student selects 2 samples "accept" and "reject", which start with the student's variant.
Variant for HW#3 must be taken the same as for HW#1

The completed task must be sent in the following form:

1) File/scripts with research (there must be comments, without comments the task is considered unresolved)
2) Excel file with answers to the following questions:

1. What is the proportion of 1 in the "accept" sample?
2. It is necessary to calculate the following for all interval variables:
 - Proportion of missing values
 - Median
 - The average
 - Standard deviation
 - Information Value for each variable
3. It is necessary to calculate for all categorical variables the following:
 - Fashion
 - Proportion of missing values
 - Information Value
 - Are there outliers, abnormal values
4. Build logistic regression only on approved applications with transformed WoE variables. What is the meaning of GINI? F1 measure?
5. Conduct a Reject Inference analysis. What is the proportion of rejected applications out of the entire sample of applications?
6. Build a logistic regression on all claims with transformed WoE variables. What is the meaning of GINI, F1? Has the model changed?

7. What model do you recommend for implementation in a production environment? Give a detailed explanation

Assessment for homework No. 3 is set on a 2-point scale, where "2" - the task is solved completely, "1" - the task is not completely solved or with shortcomings, "0" - the task is not solved or solved incorrectly.
The transfer of grades for homework from a 2-point scale to a 10-point scale is carried out by multiplying the grade by 5 without rounding.

Homework #3 will be graded:
"2" - the model is correctly built both on the accept sample and on the reject sample. Correct answers given.
"1.6" - the model is correctly built both on the accept sample and on the reject sample. 50% of answers are correct.
"1.4" - the model is correctly built both on the accept sample and on the reject sample. Wrong answers given
"1" - the model is built only on the accept sample.
"0.8" - the task is not completely solved. 50% of answers are correct.
"0" - the task is not solved or solved incorrectly.

Deadline - until 28.03.2022 00-00 (Moscow time) inclusive

Solutions should be sent to Maria.Vorobyova.Ser@gmail.com and msvorobeva@hse.ru
The subject of the letter must contain the following: HSE + Course number + variant_number + full name.

The file name must include:
*Course number
*Option number
*FULL NAME

Example: "PMI_3course_Option_8_IvanovIvanIvanovich"

If the works will repeat each other, both works will be considered unresolved.