Домашнее задание по курсу «Машинное обучение», тема «Прогнозирование»

Часть I

1. В файле «domations.csv» находится выборка с информацией об участниках программы пожертвования денег на нужды ветеранских организаций. Каждая запись – один человек из списка рассылки. У него есть социо-демографические признаки (пол, возраст, медианная оценка дохода в районе его проживания, является ли он домовладельцем и т.д. и т.д.), поведенческие признаки (агрегированные характеристики его ранних пожертвований типа GiftCount36 – число пожертвований за три года, GiftAmntLast – сумма последнего пожертвования, PromCntCard12 – число контактов с ним за год в рамках рекламной компании и т.д.). Также есть два отклика: флаг TargetB (пожертвовал или нет) и TargetD – сумма пожертвования (пропуск, если не жертвовал, иначе сумма в долларах). В рамках первой части задания нужно построить регрессионные модели (только по людям, кто пожертвовал деньги), объясняющие и прогнозирующие сумму пожертвования TargetD.

Name	Type	Label
TargetB	Numeric	Target Gift Flag
ID	Character	Control Number
TargetD	Currency	Target Gift Amount
GiftCnt36	Numeric	Gift Count 36 Months
GiftCntAll	Numeric	Gift Count All Months
GiftCntCard36	Numeric	Gift Count Card 36 Months
GiftCntCardAll	Numeric	Gift Count Card All Months
GiftAvgLast	Currency	Gift Amount Last
GiftAvg36	Currency	Gift Amount Average 36 Months
GiftAvgAll	Currency	Gift Amount Average All Months
GiftAvgCard36	Currency	Gift Amount Average Card 36 onths
GiftTimeLast	Numeric	Time Since Last Gift
GiftTimeFirst	Numeric	Time Since First Gift
PromCnt12	Numeric	Promotion Count 12 Months
PromCnt36	Numeric	Promotion Count 36 Months
PromCntAll	Numeric	Promotion Count All Months
PromCntCard12	Numeric	Promotion Count Card 12 Months
PromCntCard36	Numeric	Promotion Count Card 36 Months
PromCntCardAll	Numeric	Promotion Count Card All Months
StatusCat96NK	Character	Status Category 96NK
StatusCatStarAll	Numeric	Status Category Star All Months
DemCluster	Character	Demographic Cluster
DemAge	Numeric	Age
DemGender	Character	Gender
DemHomeOwner	Character	Home Owner
DemMedHomeValue	Currency	Median Home Value Region
DemPctVeterans	Numeric	Percent Veterans Region
DemMedIncome	Currency	Median Income Region

2. Выберите и сохраните в качестве проверочной выборки (holdout) 30% исходной выборки со стратификацией по отклику. Обратите внимание, что отклик непрерывный и его нужно

дискретизировать. Число интервалов и метод дискретизации выберите самостоятельно. Постройте и визуализируйте гистограмму (или kde апроксимацию) для распределения отклика во всем исходном наборе, в проверочной и в тренировочной выборках.

3. На этапе предобработки данных сделайте подстановку пропусков методом из вашего варианта с сохранением бинарных признаков о том, какие переменные были проимпутированы. Преобразования категориальных переменных с помощью WOE, Target encoding, Threshold encoding и других методов, а также преобразование числовых переменных (для получения более симметричных распределений с помощью log или Box-Cox) приветствуется, но не обязательно.
4. Произведите отбор важных переменных с помощью линейного регрессионного метода из вашего варианта, перебрав все возможные сложности моделей в рамках вашего метода и выбрав лучшую по кросс-валидации с 5 блоками и MSE в качестве критерия. В пошаговых регрессионных методах для остановки и выбора следующего шага используйте R-квадрат, p-value или AIC на ваше усмотрение. Постройте график зависимости CV-MSE от сложности (число переменных или число компонент в модели), график трассы стандартизованных коэффициентов от сложности. Вертикальной линией на этих графиках обозначьте лучшую по CV сложность модели.
5. Для лучшей выбранной сложности линейной модели с помощью бутстреппинга (100 бутстреп выборок размера 25% от исходной) постройте гистограммы (или kde аппроксимацию) распределения константы смещения в полученном регрессионном уравнении (константы b если регрессии y=ax+b) с указанием на графике среднего значения и 95% интервала. Аналогично оцените OOB ошибку MSE. Как она соотносится с лучшей кросс-валидационной ошибкой и ошибкой на проверочной части выборки?
6. Используйте отобранные переменные для построения нелинейной модели прогнозирования числового отклика с помощью метода из вашего варианта, при этом отбирая метапараметры также с помощью метода из вашего варианта. Замечания:
    - В PLS регрессиях для отбора переменных (после отбора числа компонент по кросс-валидации) используйте VIP статистику с любым порогом в диапазоне [0.5,1].
    - Обратите внимание, что категориальные переменные можно либо включить в модель целиком (со всеми уровнями), либо не включать.
    - Для однослойного MLP можно варьировать число нейронов и константу регуляризации, для Poisson Regression, Gamma Regression и полиномиальной гребневой регрессии - константу регуляризации и степень полинома (для Gamma и Poisson воспользуйтесь PolynomialFeatures).
7. Постройте график – «решетку» перебора метапараметров, цветом указав качество моделей, а размером точек – число повторов для halving). Сравните CV, OOB и holdout оценки качества полученных линейных и нелинейных моделей, какие выводы из этого можно сделать?

Запишите и перешлите для проверки JN реализующий шаги 1-7.

ВАРИАНТ	ПУНКТ 3	ПУНКТ 4	ПУНКТ 6
1	KnnImputer (neighbors=3)	Forward OLS	MLP (tanh), HalvingGridSearchCV
2	KnnImputer (neighbors=5)	Backward OLS	MLP (relu) , HalvingGridSearchCV
3	KnnImputer (neighbors=7)	LASSO_LARS	Poisson Regression, HalvingGridSearchCV
4	SimpleImputer (median)	LARS	Gamma Regression, HalvingGridSearchCV
5	SimpleImputer (mean)	PLS	Poly Regression, HalvingGridSearchCV
6	KnnImputer (neighbors=3)	PLS	Poisson Regression, HalvingRandomSearchCV
7	KnnImputer (neighbors=5)	LARS	Gamma Regression, HalvingRandomSearchCV
8	KnnImputer (neighbors=7)	LASSO_LARS	Poly Regression, HalvingRandomSearchCV
9	SimpleImputer (median)	Backward OLS	MLP (tanh) , HalvingRandomSearchCV
0	SimpleImputer (mean)	Forward OLS	MLP (relu) , HalvingRandomSearchCV

Часть II.

«Творческое задание» на классификацию. Целью задания является освоение алгоритмов и методов прогнозирования для решения учебной задачи анализа данных в условиях, близких к реальным условиям, возникающим при решении прикладных задач анализа данных. Дан тот же набор данных с известным откликом, а также тестовый набор “test.csv”, где реальный отклик не будет известен вам, но будет известен проверяющему. Необходимо построить модель, которая выберет среди людей из набора test.csv тех, кому имеет смысл делать предложение о пожертвовании при условии, что стоимость контакта с каждым человеком фиксированная и равна 0.68 USD. Таким образом вы можете использовать три различных подхода:

- С помощью модели прогнозировать сумму пожертвования P_TargetD (с учетом того, что есть много нулей) и сформировать список ID, кому стоит делать предложение, у кого ожидаемая сумма P_TargetD> 0.68.
- С помощью нелинейной модели прогнозировать вероятность положительного отклика P_TargetB и считать ожидаемую сумму пожертвования (риск) как P_TargetB *E(TargetD)-0.68*(1- P_TargetB), выбирать тех, у кого ожидаемая сумма (положительный риск) будет больше нуля. В качестве средней оценки E(TargetD) можно брать мат. ожидание, медиану или другую оценку центра масс или моды распределения.
- С помощью двух моделей или с помощью одной модели (например, нейросети с двумя выходами): одна для прогноза суммы P_TargetD, вторая для прогноза вероятности пожертвования P_TargetB, посчитать тот же риск, что и в варианте выше: P_TargetB *P_TargetD-0.68*(1- P_TargetB) и выбрать по нему.

Третий вариант самый правильный, но самый сложный. Оцениваться качество модели будет так:

- На тестовом наборе, где реальный отклик не будет известен студенту (вам), но будет известен проверяющему (мне) нужно применить свою модель (свои модели) и переслать вместе с JN csv файл с одной колонкой отобранных ID из тестового набора.
- С учетом того, что в тестовом наборе у проверяющего есть информация о факте и сумме пожертвования в качестве оценки будет использоваться общая сумма, собранная по отобранным клиентам с учетом стоимости контакта.
- Для сдачи задания необходимо получить сумму на тестовом наборе больше $11K (у случая «выбрать всех без моделирования» сумма где-то $10K). Набравшие больше $14K получают экзамен автоматом (если смогут объяснить как строится и работает модель) вне зависимости от посещаемости и сдачи других заданий.

При решении задачи можно использовать любые открытые общедоступные на python пакеты и алгоритмы, рассмотренные в курсе, включая sklearn, pytourch, keras и другие. При использовании алгоритмов, не рассмотренных в курсе, нужно быть готовым подробно рассказать про построенную модель и методы построения.