Задание 3. «Прогнозирование. kNN и регрессии»

1) Запустите SAS Enterprise Miner, откройте проект с именем “Prak”.
2) В дереве проекта создайте новую диаграмму с именем Diag3
3) На диаграмму добавьте узел Import Data из раздела Sample и подключите csv набор данных white_wine. В наборе указана целевая (target) переменная target_quality типа Interval, которая описывает субъективную оценку качества сорта вина, заданную экспертом. Остальные переменные - входные (все Interval). Задача – спрогнозировать оценку эксперта в зависимости от объективных химических показателей вина и вявить важные признаки.
4) После узла Import Data подключите узел Transform Variables и установите дискретизацию числового отклика на квантили: Interval Target = Quantile. Далее из раздела Utility подключите узел Metadata (позволяет менять метаданные внутри процесса анализа). Выберети раздел variables->train и для переменной с префиксом PCTL_ поставьте новую роль Rejected (чтобы работать только с исходным откликом). Затем подключите узел Data partition, оставьте разбиение по умолчанию 40 – train и по 30 на тетсовый и валидационный наборы. Откройте список переменных на этом узле и поставьте роль для переменной с префиксом PCTL_ Stratification. Зачем это делается?
5) Подключите узел Transform variables после Data Partition и сделайте настройку в разделе Interval Inputs – Standardize и подключите MBR после transform variables. Укажите в качестве числа соседей Number of Neighbors стандартную эвристику примерно sqrt(N), задайте 20. После MBR подключите узел Model Comparison. Сделайте в нем настройку в разеделе Model selection->Selection Statistics->Average Squared Error, Selection Table - > Test. Какие значения ASE получились на тренировочном, валидационном и тестовом наборе данных?
6) После узла Transform variables и перед узлом MBR вставьте узел PCA (Вариант I), Variable clustering (Вариант II). Сколько переменных теперь передается на вход MBR? Как изменились значения ASE в результате? Почему это могло произойти?
7) Параллельно от узла Data Partition подключите узел Regression и сделайте следующие натсройки. В разделе Model selection задаейте метод Selection Model Backward (Вариант I) и Forward (Вариант II). Критерий для выбора модели Selection Criterion поставьте Validation Error. Выход узла регрессион соедините с узлом сравнения моделей. Какая модель показывает лучшее качество на тестовом наборе? Какие переменные не вошли в результирующую регрессионую модель? На каком шаге была выбрана лучшая модель (см. график Iteration plot (View->Model)? Посмотрите в журнале (раздел Output) какая из вошедших переменных наименее важная с точки зрения t-статистики?
8) Добавьте узел Transform variables перед узлом Regression и сделайте настройку Interval Inputs – Maximum Normal. Как изменились результаты регрессионной модели с точки числа степеней свободы и значения ASE на тестовом наборе?
9) Отредактируйте уравнение регресии, чтобы оно стало полным полиномом второй степени, в разделе Equation задайте Polynomial Terms = Yes, Two-factor interactions = yes. Normal. Как изменились результаты регрессионной модели с точки числа степеней свободы и значения ASE на тестовом наборе?
10) После узла Data partition добавьте узел PLS и соедините его выход с узлом сравнения моделей. Задайте алгоритм Regression method = PLS (I Вариант), PCR (II вариант). Задайте настройки автоматического отбора числа факторов: CV Method = Test Set в Cross Validation. Сколько компонент было выбрано? Какие переменные были отобраны по критерию VIP (I Вариант), по абсолютным значениям коэфициентов (II Вариант). Какое значение ASE модель показывает на тестовом наборе?

11) После узла Data partition добавьте узел LARS и соедините его выход с узлом сравнения моделей. Задайте алгоритм отбора переменных LASSO и критерий для выбора моделей AIC (I вариант) или SBC (II Вариант). Какое значение ASE модель показывает на тестовом наборе? Какая пременная была отобрана на первом шаге? Сколько всего переменных было отобрано? Обратите внимание на компоненту Iteration Plot в результатах. Если бы вместо критерия отбора модели вашего варианта использовался критерий ASE на валидационном наборе, то было бы отобрано больше или меньше переменны?
12) Ответы на вопросы включите в текстовый файл (в pdf формате), запишите диаграмму в формате xml (нажав на диаграмму в дереве проекта и выбрав Save as). Перешлите текстовый файл с ответами и xml с диаграммой в качестве результата зада