КУРСОВАЯ РАБОТА

Аналитический отчет по дисциплине «Классическое машинное обучение»

Выполнил:

Студент группы М24-525

Воронин Михаил Михайлович

Цель исследования:

Разработка и сравнение различных моделей ML для:

- Регрессия:
  - IC50
  - CC50
  - SI
- Классификация:
  - превышает ли значение IC50 медианное значение выборки
  - превышает ли значение CC50 медианное значение выборки
  - превышает ли значение SI медианное значение выборки
  - превышает ли значение SI значение 8

Этапы предобработки:

- Удаление столбцов с нулевым стандартным отклонением
- Удаление технического столбца (unnamed: 0)
- Замена null значений на медианные значения по этому столбцу
- Удаление строк, которые являются выбросами. То есть выше 99 перцентиля
- PCA для сокращение размерности, но позже была удалена, т.к. данные нелинейно зависимы
- Масштабирование признаков с помощью StandardScaler
- Логарифмирование целевых признаков для сглаживания выбросов
- Отбор 50 наиболее значимых признаков по отношению к каждой целевой переменной с помощью библиотеки sklearn.feature_selection
- Для каждой модели были подобраны лучшие гиперпараметры с помощью GridSearch

Задачи регрессии:

Были использованы 5 моделей для поиска регрессии, такие как:

- LinearRegression
- Ridge
- Lasso
- RandomForest
- XGBoost

IC50	
Модель	R^2
LinearRegression	0,39
Ridge	0,34
Lasso	0,33
RandomForest	0,61
XGboost	0,63

CC50	
Модель	R^2
LinearRegression	0,31
Ridge	0,31
Lasso	0,31
RandomForest	0,48
XGboost	0,50

SI	
Модель	R^2
LinearRegression	0,17
Ridge	0,16
Lasso	0,15
RandomForest	0,41
XGboost	0,39

Задачи классификации:

Были использованы 4 модели для классификации данных, такие как:

- Logistic Regression
- RandomForest
- XGBoost
- KNN

IC50				
Модель	Accuracy	Precision	Recall	F1
Logistic Regression	0,72	0,80	0,70	0,75
RandomForest	0,76	0,80	0,70	0,75
XGBoost	0,76	0,80	0,70	0,75
KNN	0,74	0,75	0.72	0.73

CC50				
Модель	Accuracy	Precision	Recall	F1
Logistic Regression	0,82	0,77	0,84	0,80
RandomForest	0,82	0,78	0,85	0,81
XGBoost	0,87	0,84	0,86	0,85
KNN	0,75	0,72	0,72	0,72

SI > median				
Модель	Accuracy	Precision	Recall	F1
Logistic Regression	0,65	0,61	0,65	0,63
RandomForest	0,65	0,61	0,64	0,62
XGBoost	0,63	0,58	0,66	0,62
KNN	0,64	0,58	0,73	0,65

SI > 8				
Модель	Accuracy	Precision	Recall	F1
Logistic Regression	0.78	0,64	0,74	0,69
RandomForest	0,72	0,63	0,41	0,50
XGBoost	0,72	0.59	0,5	0,54
KNN	0,69	0,54	0,48	0,51

Заключение:

- XGBoost показал высокую эффективность в задачах регрессии и большинстве задач классификации, что делает его лучшим выбором
- Для задачи Si > 8, лучше взять обычную логистическую регрессию
- Рекомендуется логарифмировать все целевые значения перед обучения, чтобы стабилизировать обучение и повысить точность
- В задачах классификации можно использовать ансамбль моделей, например XGBoost и Logistic Regression, чтобы повысить устойчивость предсказаний
- Несмотря на высокую точность XGBoost, в задачах где критично интерпретируемость, стоит взять logistic regression и RandomForest