УДК 004.89

МЕТОДЫ ИНТЕРПРЕТАЦИИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Хакимов Р.А. (Университет ИТМО)

Научный руководитель – д.т.н., старший научный сотрудник Котельников Е.В.

(Национальный центр когнитивных разработок Университета ИТМО)

В ходе данной работы рассмотрены вопросы значимости интерпретации моделей, последние достижения в интерпретируемости моделей машинного обучения, затронуты актуальные методы интерпретации, проанализированы основные принципы методов глобальной и локальной интерпретации.

Введение.

Методы машинного обучения используются для принятия решений, влияющих на отдельных людей, и, как следствие, вызывают много вопросов со стороны общества и исследователей. Можно ли доверять полученной модели? Корректна ли такая модель? Какие факторы наиболее важны? Не получилось ли решение дискриминационным? Что можно сделать, чтобы изменить результат?

Основная часть.

Основными причинами необходимости интерпретации ML-моделей выступают юридические и этические аспекты, факторы доверия к работе модели, а также потребность в тестировании и улучшении модели. Прикладное применение методов интерпретации моделей машинного обучения наиболее востребовано в тех областях принятия решений, в которых стоимость ошибки чрезвычайно высока, например в здравоохранении и правосудии.

Способы интерпретации моделей машинного обучения делятся на локальные и глобальные. Глобальные методы призваны показать, какие факторы в целом оказывают наибольшее влияние на структуру модели и на ее предсказания. Локальные методы пытаются объяснить то, как было сделано данное конкретное предсказание. Зачастую локальные методы могут быть использованы как основа для более глобальной интерпретации.

Выводы.

В ходе работы были рассмотрены глобальные методы интерпретации моделей – важность признаков, график частичной зависимости и визуальное исследование переменных, а также локальные методы – локальные суррогатные модели и интерпретация значения Шепли.

Экспериментальное исследование было посвящено сравнению методов интерпретации для модели прогнозирования аренды велосипедов. Были исследованы методы важности каждой функции при прогнозировании аренды велосипедов с помощью метода опорных векторов, график частичной зависимости и визуального исследования переменных. объяснения локальных суррогатных моделей для двух экземпляров набора данных о прокате велосипедов и объяснения Шепли для предсказания аренды велосипедов на один день.

Хакимов Р.А. (автор) Подпись

Котельников Е.В. (научный руководитель) Подпись