Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
Факультет: Робототехника и комплексная автоматизация
Кафедра: Системы автоматизированного проектирования

Интерпретация аппроксимирующих моделей машинного обучения

Выполнил: Конов А.В.
Консультант: к.т.н., доцент, Агасиев Т.А.
Научный руководитель: д.ф-м.н., профессор, Карпенко А.П.

Введение

Интерпретация аппроксимирующих моделей машинного обучения - это процесс объяснения того, как модель делает свои предсказания. Интерпретация включает анализ влияния признаков на предсказания модели и выявление зависимостей, а также предоставление понятных объяснений для решений, принимаемых моделью. Цель интерпретации заключается в том, чтобы обеспечить понимание и доверие к модели, а также помочь выявить причинно-следственные связи и использовать модель для принятия решений.

Цели и задачи

Цель работы – разработка ПО для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения на основе существующих методов интерпретации.

Задачи:

- Изучить аппроксимирующие модели машинного обучения;
- Изучить методы интерпретации аппроксимирующих моделей машинного обучения;
- Предложить алгоритм для автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения и реализовать его;
- Исследовать эффективность работы реализованного ПО.

Постановка задачи машинного обучения

В общем виде задача машинного обучения выглядит следующим образом:

$$f^* = \arg\min_f L(f, D), \quad (1)$$

где $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ - обучающая выборка; x_i - вектор признаков; y_i - целевая переменная для соответствующего вектора признаков; f - модель, аппроксимирующая зависимость между признаками и целевой переменной, L - функция потерь, оценивающая расхождение между предсказанными и фактическими значениями целевой переменной на обучающей выборке; f^* - оптимальная модель

Объект – это отдельный пример данных из набора данных.

Признак - это атрибут объекта, который используется для описания объекта в виде числовых или категориальных значений, которые затем используются для обучения модели.

Целевая переменная - это переменная, значение которой модель пытается предсказать на основе входных данных.

Набор данных – совокупность всех объектов.

Обучающая выборка - это часть набора данных, которая используется для обучения модели машинного обучения.

Тестовая выборка - это часть набора данных, которая используется для оценки производительности модели после её обучения на обучающей выборке.

id	пол	возраст	образование	здоровье	доход
1	М	12	Неполное среднее	Отл	0
2	М	25	Высшее	Отл	100.000
3	Ж	80	Высшее	Неуд	15.000
4	М	75	Среднее	Удов	18.000
5	Ж	27	Среднее	Отл	50.000
6	Ж	8	Неполное среднее	Хор	0
7	М	45	Среднее специал.	Удов	70.000

Модели машинного обучения

Классические:

- Линейная регрессия
- Логистическая регрессия
- Метод опорных векторов
- Наивный байесовский классификатор

На основе решающих деревьев:

- Решающие деревья
- Случайный лес
- Бустинг

Нейронные сети:

- Прямого распространения
- Свёрточные
- Реккурентные

Модели машинного обучения. Решающее дерево

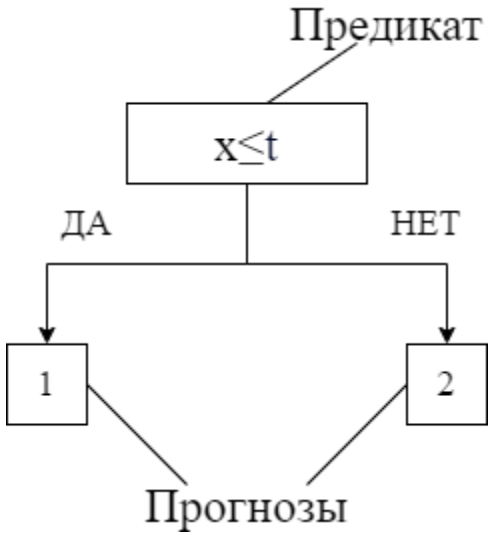
Определение решающего дерева:

- каждой внутренней вершине v приписан предикат B_v;
- каждой листовой вершине v приписан прогноз $c_v \in Y$, где Y — область значений целевой переменной.

Предикат B_v может иметь, произвольную структуру, но, как правило, на практике используют сравнение с пороговым значением $t \in R$ по произвольному j-му признаку:

$$B_v(x, j, t) = [x_j \leq t]. \quad (2)$$

При проходе через узел дерева с данным предикатом объекты будут отправлены в правое поддерево, если значение j-го признака у них меньше либо равно t, и в левое — если больше.



Методы интерпретации моделей машинного обучения

Локальные:

- LIME
- SHAP
- ICE

Глобальные:

- PDP
- ALE

Методы интерпретации моделей машинного обучения. Метод LIME

Суть метода LIME заключается в создании локальной модели, которая объясняет прогноз модели на отдельном объекте.

Математическое описание метода LIME:

$$E(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g). \quad (3)$$

Модель объяснения для объекта x – это локальная модель g, которая минимизирует функцию потерь L, которая измеряет, насколько близким является объяснение к прогнозу исходной модели машинного обучения f, при этом сохраняя низкую сложность модели $\Omega(g)$. G - это семейство возможных локальных моделей. Мера близости π_x определяет размер окрестности вокруг объекта x, которая рассматривается для объяснения.

Методы интерпретации моделей машинного обучения.

Метод SHAP

Суть метода SHAP заключается в вычислении значений Шепли для каждого признака и объединении их в одну величину, которая показывает важность признака для прогноза модели на конкретном объекте.

Математическое описание метода SHAP:

$$\Delta_f(i, S) = f_{S \cup i}(x_{S \cup i}) - f_S(x_S), \qquad (4)$$

здесь $x_{S \cup i}$ - признак, для которого рассчитываются значения SHAP; x_S - все остальные признаки.

Методы интерпретации моделей машинного обучения. Метод PDP

В основе метода PDP лежит идея оценки среднего значения прогноза модели для всех объектов, при фиксированных значениях определенных признаков, варьируя значения всех остальных признаков.

Математическое описание метода PDP:

Частичная функция f_S оценивается путем расчета средних значений на обучающей выборке, также известный как метод Монте-Карло:

$$f_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)}), \tag{5}$$

здесь x_S представляют собой признаки, для которых должна быть построена частичная функция зависимости, $x_C^{(i)}$ представляют собой значения признаков из набора данных, которые не рассматриваются, n - количество объектов в наборе данных.

Вклад определяется отклонением каждого уникального значения признака от средней кривой:

$$I(x_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(f_S\left(x_S^{(k)}\right) - \frac{1}{K}\sum_{k=1}^{K} f_S(x_S^{(k)}))}^{\,2}, \tag{6}$$

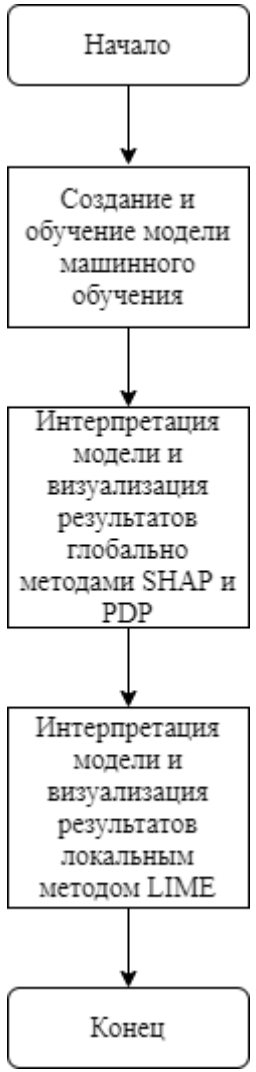
здесь $x_S^{(k)}$ представляет собой K уникальных значений признака X_S.

Предлагаемый алгоритм автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения

Алгоритм автоматизации процесса интерпретации аппроксимирующих моделей машинного обучения, включает в себя следующую функциональность:

1) создание и обучение модели машинного обучения, которую необходимо интерпретировать;
2) интерпретация модели машинного обучения различными методами;
3) визуализация результатов интерпретации модели машинного обучения.

Алгоритм в качестве входных параметров принимает набор данных, целевую переменную, номер объекта, который необходимо интерпретировать локально и тип решаемой задачи. Создается и обучается аппроксимирующая модель машинного обучения, затем модель интерпретируется глобально и визуализируется результат интерпретации, затем модель интерпретируется локально и визуализируется результат интерпретации.



Вычислительные эксперименты

Целью проведения данных вычислительных экспериментов является демонстрация работы ПО для автоматизации интерпретации методов машинного обучения для задачи классификации.
Вычислительные эксперименты проводились на существующих наборах данных:

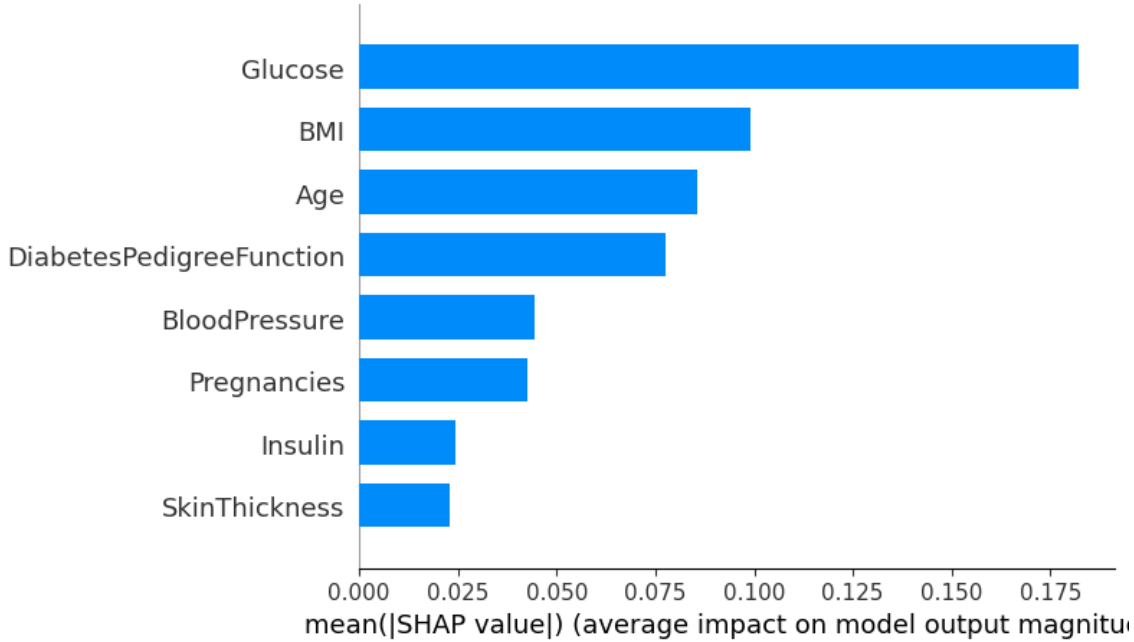
- ***Эксперимент 1***

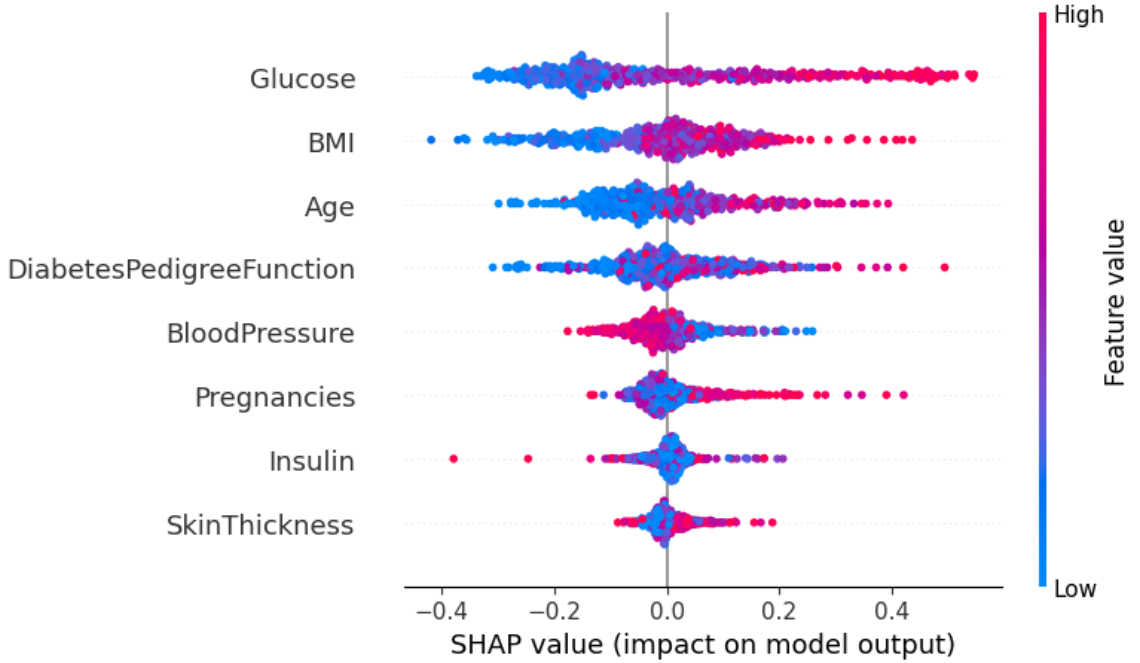
Набор данных ***diabetes***, который содержит медицинские данные, влияющие на риск развитие сахарного диабета. Объекты набора данных состоят из 8 признаков и целевой переменной: Pregnancies - количество беременностей; Glucose - плазменные концентрации глюкозы в крови; BloodPressure - диастолическое артериальное давление; SkinThickness - толщина кожи в области трицепса; Insulin - количество инсулина в крови; BMI - индекс массы тела; DiabetesPedigreeFunction - оценка предрасположенности к диабету; Age - возраст; Outcome – целевая переменная, показывающая прогноз, на предрасположенность к заболеванию сахарным диабетом в ближайшие пять лет. Набор данных состоит из 768 объектов.

- ***Эксперимент 2***

Набор данных ***california housing***, который содержит данные о средней стоимость домов в Калифорнии в зависимости от квартала. Объекты набора данных состоят из 8 признаков и целевой переменной: Longitude - долгота квартала с недвижимостью; Latitude - широта квартала с недвижимостью; HouseAge - медиана возраста домов в квартале; AveRooms - общее количество комнат в квартале; AveBedrms - общее количество спален в квартале; Population - население квартала; AveOccup - количество семей в квартале; MedInc - медианный доход в квартале; MedHouseVal - целевая переменная, показывающая медианную стоимость дома в квартале. Набор данных состоит из 20640 объектов.
Построение аппроксимирующих моделей для данных экспериментов производилось с помощью библиотеки XGBoost и дало точность для эксперимента 1 79,84%, для эксперимента 2 86,27%.

Эксперимент 1. Метод SHAP

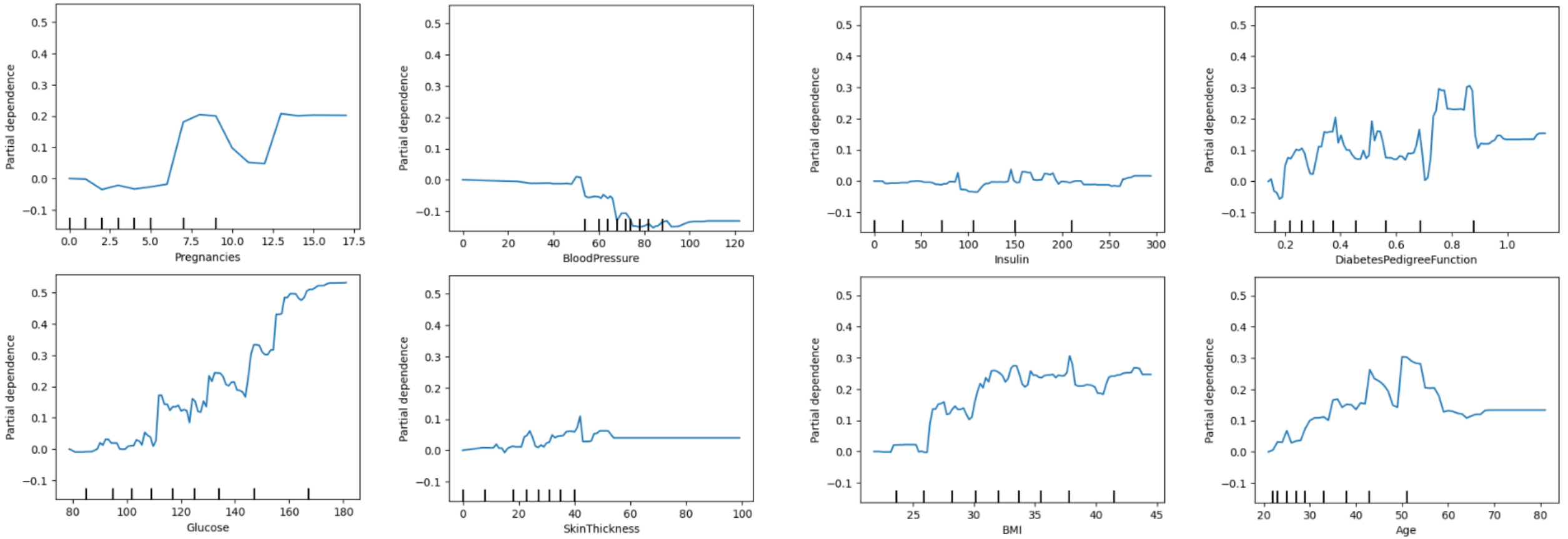




Из левого рисунка можно сделать вывод, что самыми значимыми признаками, влияющими на развитие сахарного диабета в ближайшие пять лет, являются плазменные концентрации глюкозы в крови, индекс массы тела и возраст. Наименее значимыми признаками являются количество инсулина в крови и толщина кожи в области трицепса. Признаки оценка предрасположенности к диабету, диастолическое артериальное давление, количество инсулина в крови вносят умеренный вклад в риск развития сахарного диабета в ближайшие пять лет у объекта.

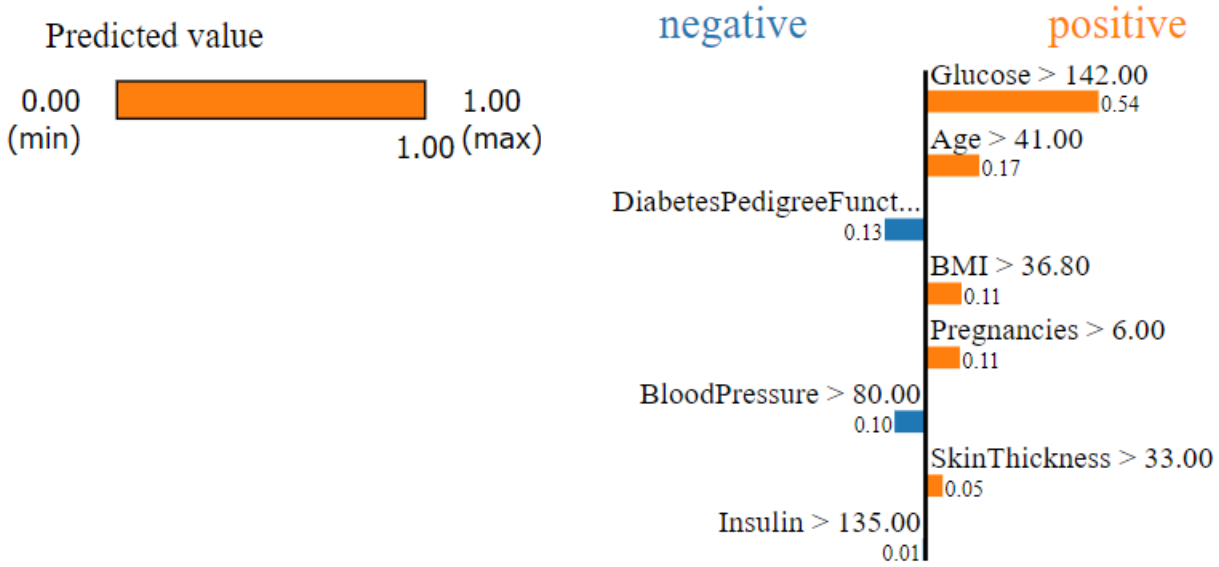
Из правого рисунка можно сделать вывод, что чем выше значение таких признаков как плазменные концентрации глюкозы в крови, индекс массы тела, возраст, оценка предрасположенности к диабету, количество беременностей и толщина кожи в области трицепса, тем выше риск развития сахарного диабета в течении пяти лет. При этом, чем выше значения признаков количество инсулина в крови и диастолическое артериальное давление, тем риск развития сахарного диабета в течении пяти лет ниже.

Эксперимент 1. Метод PDP



Из рисунка можно сделать выводы аналогичные выводам, представленным на предыдущем слайде для левого рисунка.

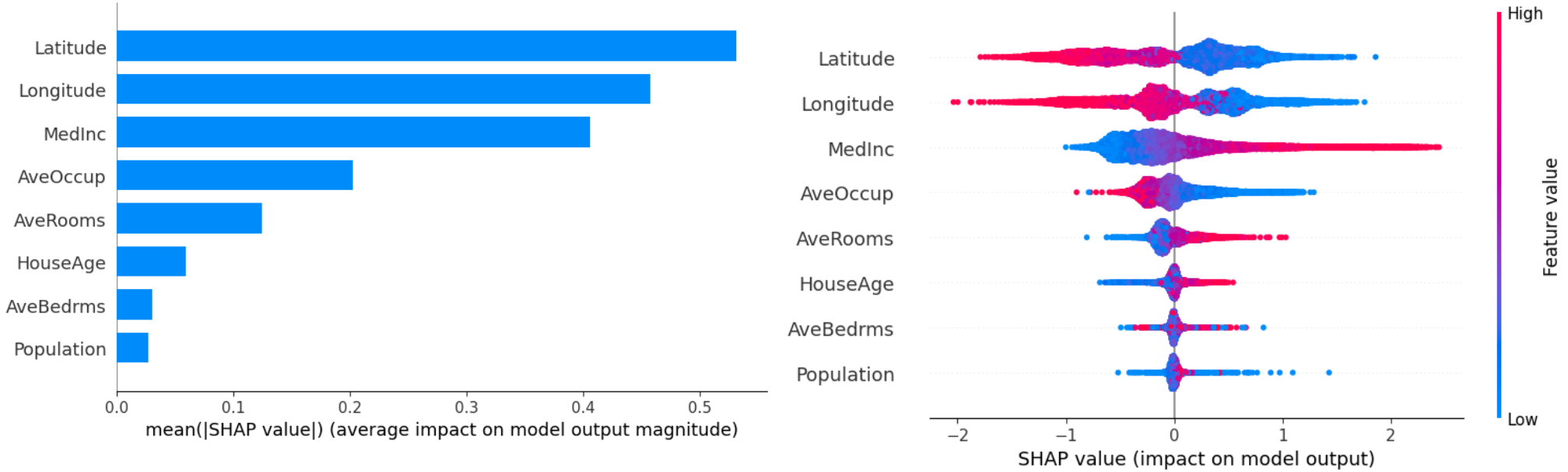
Эксперимент 1. Метод LIME



Feature	Value
Glucose	167.00
Age	43.00
DiabetesPedigreeFunction	0.17
BMI	37.60
Pregnancies	8.00
BloodPressure	106.00
SkinThickness	46.00
Insulin	231.00

Из рисунка можно сделать вывод, что несмотря на то что значения признаков оценка предрасположенности к диабету, диастолическое артериальное давление и количество инсулина в крови не приводят к предрасположенности к развитию сахарного диабета у объекта, значения остальных признаков показывают, что у объекта есть предрасположенность к развитию сахарного диабета в течении пяти лет.

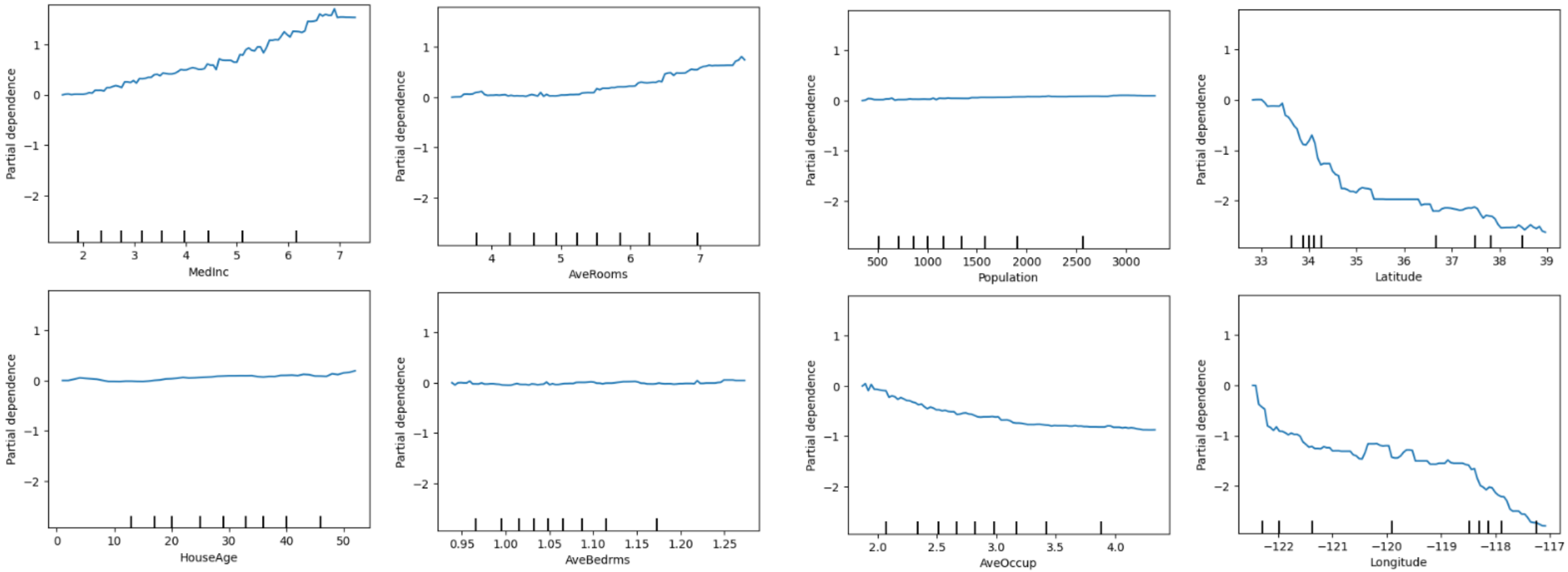
Эксперимент 2. Метод SHAP



Из левого рисунка можно сделать вывод, что самыми значимыми признаками, влияющими медианную стоимость дома в квартале, являются широта квартала с недвижимостью, долгота квартала с недвижимостью, медианный доход в квартале и количество семей в квартале. Наименее значимыми признаками являются общее количество спален в квартале и население квартала. Признаки общее количество комнат в квартале и медиана возраста домов в квартале вносят умеренный вклад в медианную стоимость дома в квартале.

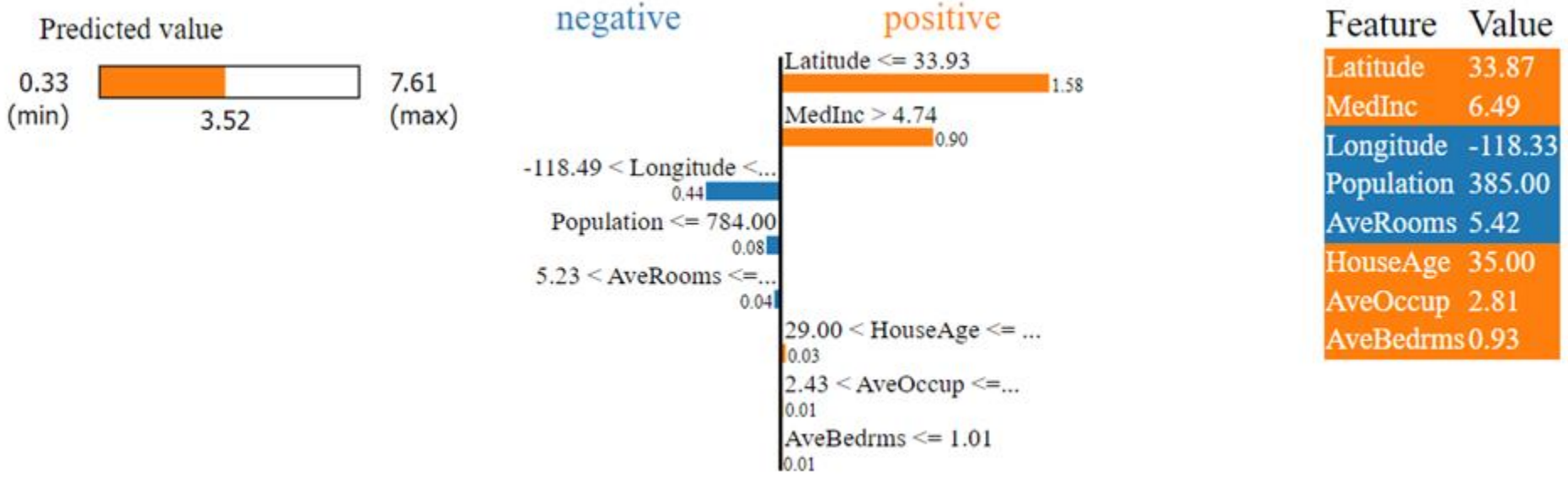
Из правого рисунка можно сделать вывод, что чем выше значение таких признаков как общее количество комнат в квартале, медиана возраста домов в квартале, медианный доход в квартале тем выше медианная стоимость дома в квартале. При этом, чем выше значения признаков широта квартала с недвижимостью, долгота квартала с недвижимостью, количество семей в квартале, тем ниже медианная стоимость дома в квартале. Значения признаков количество спален в квартале и население квартала почти не влияют на медианная стоимость дома в квартале.

Эксперимент 2. Метод PDP



Из рисунка можно сделать выводы аналогичные выводам, представленным на предыдущем слайде для левого рисунка

Эксперимент 2. Метод LIME



Из рисунка можно сделать вывод, что на повышение медианной стоимости дома в квартале повлияли все признаки кроме долготы квартала с недвижимостью, общего количество комнат в квартале и населения квартала. Они снизили медианную стоимость дома в квартале.

Заключение

В результате проведенной работы:

1) изучены аппроксимирующие модели машинного обучения;
2) изучены методы интерпретации аппроксимирующих моделей машинного обучения;
3) разработано ПО, реализующее предложенный алгоритм автоматизации процесса интерпретации;
4) проведены вычислительные эксперименты, показавшие эффективность разработанного ПО.

БЛАГОДАРЮ ЗА ВНИМАНИЕ!