**Методы исследования**

Для построения прогноза в рамках данной работы будут рассмотрены следующие модели:

- модель, основанная на градиентном бустинге,

- модель, основанная на ансамбле моделей гребневой регрессии (для прогнозирования линейной части временного ряда) и градиентном бустинге (для прогнозирования нелинейной части временного ряды).

Рассмотрим данные методы более подробно.

Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно. Основан на градиенте функции ошибки — все производные по всем значениям функции. То есть на каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Следующая модель, которая будет добавлена в ансамбль будет предсказывать эти отклонения. Таким образом, добавив предсказания нового дерева к предсказаниям обученного ансамбля мы можем уменьшить среднее отклонение модели, которое является таргетом оптимизационной задачи. Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил "ранней остановки"[10].

Модель градиентного бустинга была выбрана по следующим причинам:

- данный метод показывает лучшие предсказательные результаты на различных соревнованиях, в т.ч. Kaggle;

- методы градиентного бустинга слабо чувствительны к выбросам, поэтому можно оставить данные как есть;

- градиентный бустинг не требуют нормализации данных.

В качестве метода построения была выбрана библиотека CatBoost – библиотека градиентного бустинга, созданная Яндексом [11]. Данная библиотека показывает наилучшие результаты в сравнении с аналогами: XGBoost, LightGBM, H2O. CatBoost гораздо проще в использовании в основном за счет того, что он работает с категориальными переменными, то есть метод не требует предварительной обработки данных, которая может занять больше всего времени в типичном процессе построения модели.

CatBoost использует градиентное усиление и основан на обучение с помощью небрежных деревьев решений. По итогу модель позволяет строить симметричные забывчивые деревья, которые позволяют[12]:

1) достигнуть значительное ускорение производительности по сравнению со стандартным послойным построением деревьев;

2) уменьшить переобучения;

3) повысить устойчивость к изменению гиперпараметров.

Приведем формулы, используемые в этом алгоритме. Основное уравнение для модели градиентного бустинга [11] (1):

, (1)

где:

- – предсказание модели после M итераций;

- – базовый алгоритм (в нашем случае дерево решений);

- – коэффициент, определяющий вклад каждого дерева.

CatBoost использует слегка модифицированную версию алгоритма градиентного бустинга, который использует обучение с упорядоченными целями. Когда дерево обучается на наборе данных, оно обрабатывает объекты в случайном порядке. Когда объект предсказывается, для кодирования категориальных переменных используется среднее значение его предыдущих значений. Это делается для того, чтобы избежать утечки целевых значений, что часто случается при обработке категориальных признаков [13].

Как и в большинстве других алгоритмов, CatBoost использует функцию потерь, которая минимизируется на каждой итерации. Функция потерь выбирается в зависимости от задачи. Например, для задачи регрессии часто используется RMSE. CatBoost использует общие подходы к регуляризации, аналогичные тем, что применяются в других алгоритмах градиентного бустинга, чтобы избежать переобучения модели.

Исходя из исследований, рассмотренных в литературном обзоре, можно сделать выводы, что наилучшие результаты показывать совмещенные модели, которые включают в себя несколько алгоритмов машинного обучения.

Ансамблевый метод – метод машинного обучения, где несколько моделей обучаются для решения одной и той же проблемы и объединяются для получения лучших результатов [14].

Основная предпосылка заключается в том, что результат работы нескольких моделей будет более точен, чем результат только одной модели.

Архитектура пакетированной модели содержит две или более базовых моделей, часто называемых моделями нулевого уровня, а метамодель, которая объединяет прогнозы базовых моделей, — моделью первого уровня. Метамодель обучается на основе прогнозов, сделанных базовыми моделями на данных вне выборки [15].

Тем не менее, смешивание имеет определённые коннотации для построения модели пакетированного ансамбля. Смешивание может предложить разработку стек-ансамбля, где базовые модели — это модели машинного обучения любого типа, а метамодель — это линейная модель, которая «смешивает» прогнозы базовых моделей. Например, модель линейной регрессии при прогнозировании числового значения или модель логистической регрессии при прогнозировании метки класса вычисляют взвешенную сумму прогнозов, сделанных базовыми моделями, и будут рассматриваться как смешение прогнозов [15].

Таким образом, определим модели для построения ансамбля. Как упоминалось ранее, лучше всего использовать комбинацию методов для прогнозирования линейной и нелинейной части данных.

Для линейной части регрессии в качестве мета модели выберем модель гребневой регрессии (ридж-регрессии), так как она хороша для данных, которые демонстрируют сильную мультиколлинеарность. Она позволит снизить размерность необходимых данных и определить основные зависимости между признаками. Гребневая регрессия основана на обычной линейной регрессии, но отличается от нее большей регуляризации весов.

Для нелинейной части в качестве базовой модели будет использован уже описанный ранее алгоритм CatBoost. Таким образом, ошибки предсказания алгоритма CatBoost будут исправляться с помощью гребневой регрессии.

Оценка моделей будет производиться по следующим показателям на основе предсказания, как по обучающей выборке, так и по тестовой:

- cредняя абсолютная ошибка (MAE) (2):

, (2)

где – число примеров обучающей выборки, – целевое значение i-го примера, – предсказанное моделью значение;

- cредняя абсолютная ошибка в процентах (MAPE) (3):

, (3)

- cимметричная средняя абсолютная процентная ошибка (SMAPE) (4):

, (4)

- cреднеквадратическое отклонение (RMSE) (5):

, (5)

- коэффициент детерминации (R2) (6):

. (6)

Также модели оцениваются с помощью кросс-валидации с разделением на 5 частей. Помимо этого для интерпретации модели были рассчитаны значения Шэпли (Shapley values) для каждого признака с целью определения силы влияния на целевую переменную. Формула расчета для i-признака представлена ниже (7):

, (7)

где

– это предсказание модели с i-тым признаком,

– предсказание модели без i-того признака,

– количество признаков,

– произвольный набор признаков без i-того признака.

Подбор гиперпараметров моделей с целью их тюнинга будет осуществляться методом поиска по сетке.

Таким образом, в рамках исследования предлагается построение и улучшение двух моделей: первая – одиночная модель градиентного бустинга CatBoost, вторая – ансамблю из двух моделей (базовая модель CatBoost, метамодель – гребневая регрессия).