УДК 004.85

РАЗРАБОТКА АЛГОРИТМА ПРОГНОЗИРОВАНИЯ ИНДЕКСА ПОТРЕБИТЕЛЬСКИХ ЦЕН С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ

Калин Д.А. (ФГБОУ ВО «МГТУ «СТАНКИН»)
Научный консультант – к.т.н., доцент Ковалев И.А.
(ФГБОУ ВО «МГТУ «СТАНКИН»)

Ведение. В работе рассматривается разработка алгоритма прогнозирования индекса потребительских цен (далее – ИПЦ) в формате «месяц к месяцу». Такой вид ИПЦ отражает то, как выросла цена за месяц на основе анализа данных из Росстата, предоставленных Центральным Банком Российской федерации. Данный алгоритм может способствовать более точному прогнозированию целевого показателя ИПЦ. Были проведены успешные испытания с использованием рекуррентной нейронной сетью LTSM, показавшие работоспособность предлагаемого решения.

Основная часть. Прогнозирование индекса потребительских цен является важным инструментом для оценки уровня инфляции в экономике. В работе было предложено использовать методы машинного обучения для прогнозирования ИПЦ. В последние годы машинное обучение стало все более распространенным инструментом для прогнозирования экономических показателей, и предлагается проверить его эффективность в случае прогнозирования ИПЦ. Целью данной работы является высокая точности прогнозов и значимые показатели метрик качества модели.

Текущий набор данных включал несколько атрибутов, такие как: уникальный номер товара/услуги, дата наблюдения, Статус товара/услуги на дату наблюдения (InStock – в продаже, OutOfStock – отсутствует в продаже), Цена товара/услуги на дату наблюдения (Если StockStatus = OutOfStock – значение отсутствует). Началом временного ряда служит наблюдение с наиболее ранней датой DateObserve. Окончанием временного ряда служит наблюдение с наиболее поздней датой DateObserve и статусом товара StockStatus = OutOfStock. Если наиболее позднее наблюдение имеет статус товара StockStatus = InStock, это значит, что товар на настоящий момент в продаже по последней известной цене.

На начальном этапе была проведена масштабная работа с данными. Набор данных предполагает данные об очень большом количестве товаров и услуг (более 12 млн. ед.) и необходимо было предусмотреть в модели блок, который оптимизировал их количество. На начальном этапе были убраны дубликаты. Далее были шаги по исключению выбросов в данных, например, это были товары с очень высокой/низкой стоимостью и товары с маленькой историчностью (менее 110 записей). Так же были разрывы во временных рядах большинства товаров. Разрывы были заменены усредненными значениями всех существующих записей. При анализе временного ряда необходимо было сделать проверку на стационарность ряда, но при таких компонентах ряда, как тренд или сезонность, ряд не является стационарным. При использовании машинного обучения достаточно слабой стационарности ряда. При проверке на стационарность была выявлена слабая стационарность, что позволяло не приводить ряд к стационарному.

Для более точного прогнозирования ИПЦ на обезличенных данных, была выдвинута гипотеза о том, что дешевые товары могут сильнее влиять на индекс потребительских цен, так как они являются более важными в составлении базовой корзины товаров и услуг, используемой для вычисления индекса. Они также могут быть более доступными для большинства населения и, следовательно, иметь больший вес в индексе. Изменение цен на дешевые товары может иметь больший влияние на индекс потребительских цен, чем изменение цен на дорогие товары. Поэтому было проведено логарифмирование ряда, что также помогло привести ряд к нормальному распределению. А как известно, некоторые алгоритмы машинного обучения, в том числе нейронные сети, чувствительны к распределению данных и могут давать неоптимальные результаты, если данные имеют

смещенное или экстремальное распределение. Приведение данных к нормальному распределению помогает улучшить качество обучения и устранить неоднозначности, которые могут возникнуть при работе с данными с ненормальным распределением. Например, если данные имеют экстремальные значения, то они могут перебить все другие значения и сделать обучение неэффективным.

После получения нормализованных данных, была проведена агрегация по дате наблюдения и выбрано среднее цене товара. Расчет ИПЦ проводился по стандартной формуле:
ИПЦ товара = Pt/ Pt-1 – 1,
где Pt – цена на товар/услугу на конец месяца t Pt-1 - цена на товар/услугу на конец месяца t-1, т.к. при тестировании более сложных видов расчётов точность прогнозирования значительно падала.

При выборе модели для расчета индекса потребительских цен важно выдвигать требования к адаптивности, быстрой обучаемости и точности прогноза, поскольку это обеспечит высокую эффективность и точность модели в решении реальных задач. Адаптивность модели важна, чтобы она могла адекватно обрабатывать изменения в данных и давать адекватные прогнозы в условиях неопределенности и переменных условий. Быстрая обучаемость модели важна, чтобы она могла быстро адаптироваться к новым данным и обеспечивать точные прогнозы. В общем, выдвигая требования к адаптивности, быстрой обучаемости и точности прогноза, обеспечивалось, что модель, что выбираемая модель будет эффективно работать и предоставлять нужную информацию для принятия решений. Long Short-Term Memory (LSTM) является вариантом рекуррентных нейронных сетей, которые специально разработаны для решения задач анализа временных рядов. LSTM имеет уникальную архитектуру, которая позволяет ему лучше выучить и запомнить длительные зависимости в данных, чем другие типы рекуррентных нейронных сетей. Поэтому, если выдвигаемые требования к модели включают в себя требование к адаптивности и быстрой обучаемости, а также возможность лучше выучить и запомнить длительные зависимости в данных, то LSTM может быть идеальным выбором для прогнозирования индекса потребительских цен.

Заключение. За целевую метрику качества оценки модели была взята Mean Absoulute Error (MAE), по следующим причинам:
Интерпретация: MAE является простым и легко интерпретируемым метрикой, поскольку она представляет среднее абсолютное значение ошибок прогноза.
Равномерное распределение: MAE учитывает все ошибки прогноза без предпочтения больших или меньших ошибок, что важно при прогнозировании ИПЦ, так как этот индекс является важным показателем для макроэкономики.
Непрерывность: MAE является непрерывным метрикой, что важно для моделей машинного обучения, так как они чувствительны к изменениям в метриках.

При валидации модели метрика MAE была равна 0.0245, что является низким показателем, что в свою очередь говорит о высокой точности прогноза. Так же была получена высокая скорость обучения. При обработке датасета размером 3 Гб, время за которую данные проходят предобработку и обучение составили примерно 40-45 минут.

Список использованных источников:
1.Машинное обучение:алгоритмы для бизнеса. де Прадо Маркос Лопез,«Питер», 2019 г.,432 с.
2.Big data analytics of the technological equipment based on Data Lake architecture, Kovalev, Ilya., Nezhmetdinov, Ramil., Kvashnin, Denis, International conference on modern trends in manufacturing technologies and equipment. (ICMTMTE 2019), 2019

Калин Д.А. (автор)

Ковалев И.А. (научный консультант)