Аналитическая записка по проекту:

«Обработка данных для расчета ИПЦ в России с применением «больших данных»

Индекс потребительских цен (ИПЦ) – один из важнейших макроэкономических показателей, который влияет на поведение как частных экономических агентов – фирм и потребителей, так и ключевых регуляторов национальной экономики - Правительства и, в первую очередь, Центрального Банка.

Традиционная методология расчета ИПЦ предполагает ручной сбор ценовых данных по различным категориям товаров, агрегацию и фиксацию данных в определенные дни недели/месяца и т. д. При этом с развитием интернет-торговли и алгоритмов машинной обработки больших массивов данных появляются альтернативные способы фиксации ценовых котировок практически в любой момент времени. Такие подходы предполагают, например, автоматизированную обработку (парсинг) веб-сайтов интернет-магазинов, в результате чего становятся доступными микроданные о ценах тысяч отдельных товаров.

Работа с микроданными по потребительским ценам является основополагающим аспектом данного проекта. Анализ подобных микроданных помогает отслеживать потенциальную инфляционную ситуацию в стране, то есть имеет ли место рост цен на категории потребительских товаров внутри экономики. Актуальность данного исследования также подтверждается получением наиболее полной выборки обрабатываемых данных, так как зачастую при работе с агрегированными показателями теряется важная для конечного результата информация и/или ее характеристики.

Деятельность внутри проекта была разделена на 3 этапа:

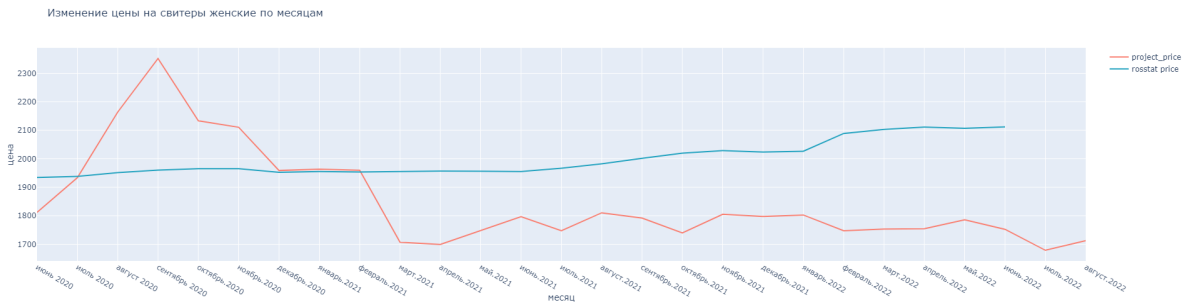
1. разметка базы данных с товарами из российских интернет-магазинов;
2. применение методов машинного обучения для классификации товаров в рубрики, выделяемые Росстатом (перед этим шагом была изучена методология ИПЦ, применяемая Росстатом при наблюдении за потребительскими ценами);
3. работа с собранными микроданными и последующая визуализация получившейся ценовой динамики.

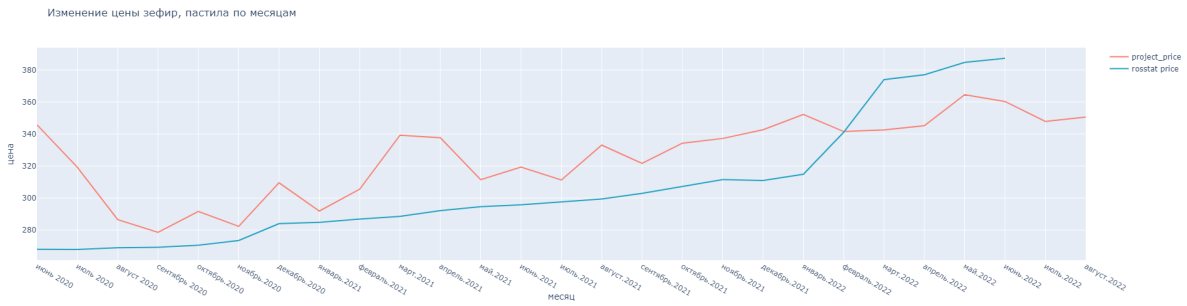
В специально разработанном графическом интерфейсе, который представлял собой локальную копию базы данных онлайн-цен, была собрана необходимая информация, относящаяся к продовольственным товарам, непродовольственным товарам и услугам. Следующий шаг подразумевал обработку получившегося датафрейма с использованием инструментария языка Python (работа с библиотеками, функциями и др.) и базовых методов машинного обучения.

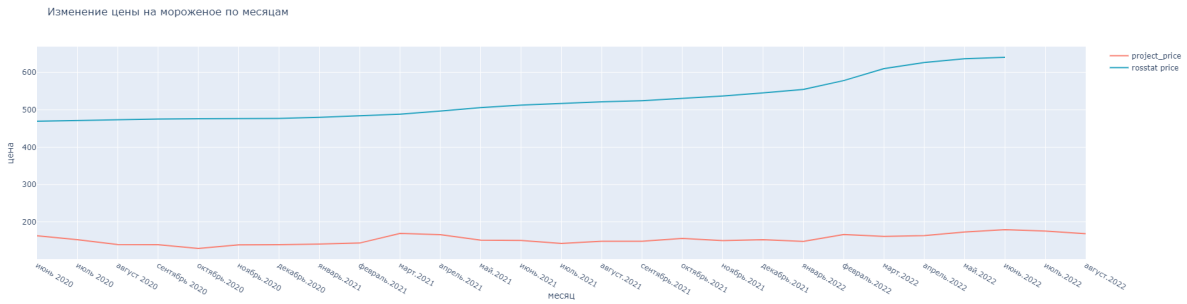
Основная задача каждого участника проекта стояла в том, чтобы, выбрав 3-5 различных категорий товаров, применить модель классификации к данным и попробовать составить «предсказание» для добавления потенциально подходящего товара в нужную выборку по определенному товару/услуге. Для успешной реализации поставленной задачи была разработана и обучена модель логистической регрессии на Python для каждой товарной категории - применен стохастический градиентный спуск с регуляризацией elastic net.

Заключительный этап проекта включал в себя создание алгоритмов фильтрации ценовых котировок, которые будут применяться для дальнейшего расчета индекса цен и решать ряд проблем, которые появляются при классификации большого массива данных (например, пропуски в наблюдениях, неправдоподобно большие/маленькие значения цен, врéменная скидка на товар/услугу). В последнее задание проекта также входило разделение товарных категорий на датафреймы и последующая аналитическая работа с каждым их них (т.е создание временного ряда, который отражает изменение цены на товары в каждый из месяцев).

Как результат, получившаяся динамика со средними ценовыми значениями была визуализирована и сопоставлена с официальной статистикой Росстата.







По представленным графикам можно понять, что данные, полученные в ходе исследования, и статистика Росстата не на всех участках таймлайна имеют схожие ценовые значения. Причин этому может быть несколько: разница в единицах измерения товарных единиц, попадание неподходящих товаров в рассматриваемые выборки и др.

Стоит отметить, что методология, предложенная данным проектом, при учете и корректировке имеющихся в ней погрешностей, потенциально может рассчитывать индекс, значение которого будет приближенным к официальному.