Пояснительная записка к курсовому проекту

Курс образовательной компании OTUS Machine Learning. Professional.

# Цель курсового проекта

Целью курсового проекта является построение модели, позволяющей на периодической основе предсказать цену продажи (закупки) на продаваемые компанией товары. Предсказания должны быть конкурентными согласно ситуации на рынке. Более амбициозная задача: *автоматическое назначение цен продажи с минимальным участием сотрудника компании*.

# Предметная область

Компания является дистрибутором оргтехники (принтеры, МФУ), расходных материалов и запчастей к ним. В номенклатуру товаров компании входит продукция таких производителей, как HP, Kyocera, OKI, Konica-Minolta, Canon. Компания импортирует или закупает большими партиями у иных импортеров оргтехнику и осуществляет дальнейшую перепродажу мелкооптовым и розничным продавцам.

Следует отметить, что один и тот же товар может быть, как импортированным, так и купленным на внутреннем рынке.

# Мотивация

Причиной, приведшей к решению о необходимости автоматического формирования цен продажи, стал обвал в объёме трафика B2B системы. Обвал был вызван неадекватными ценами продажи, что вызвало отток наших клиентов. И до этого были обвалы, но не столь грандиозные и напоминали краткосрочные колебания. Как было выяснено, неадекватные цены были результатом непонимания рыночной ситуации и заблуждений относительно того, кем, с точки зрения ведения бизнеса, являются наши клиенты и конкуренты. Ситуацию удалось выправить, но не в полном объёме. Поэтому и было принято решение использовать методы машинного обучения для формирования цен. Человек должен быть задействован только для контроля за аномалиями при автоматическом формировании цен.

Более точная формулировка: естественный интеллект некоторых сотрудников компании так достал, что приходится внедрять искусственный интеллект.

# Исходные данные

Исходными данными являются:

* цены на товары прямых конкурентов компании на начало текущего рабочего дня;
* рекомендуемая вендором розничная цена (доступна не всегда);
* дата последней продажи – количество дней с момента продажи на момент формирования предсказания;
* цена последней продажи;
* категория товара (принтер/МФУ, картридж, запасная часть);
* признак жизненного цикла товара (новый, снят с производства, регулярный);
* характеристика изменения (уменьшения) цен в зависимости от периода нахождения без движения (мало востребованный товар).

## Агрегаторы цен

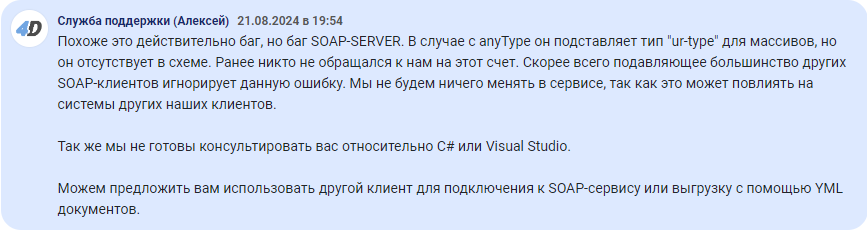
С агрегаторами ситуация, прямо скажем, нерадостная. На рынке их немного, и качество предоставляемых услуг желает лучшего. Удалось пообщаться со следующими агрегаторами:

* *4dealer.ru*;
* *price.ru*;
* *s4b.ru*;
* *proanalytics.net*;
* *tovaropediya.ru*;

У меня сложилось устойчивое мнение, что предоставлять информацию по конкретному рынку, в виде, удобном для применения в системах анализа данных, за деньги, нет даже в мыслях у агрегаторов, с которыми удалось пообщаться. Подавляющим большинством агрегаторов предполагается, что пользователь будет копошится на их сайте в поисках цен на интересующие товары.

Особо следует выделить агрегатора цен *ProAnalytics* (<https://my.proanalytics.net/>). Единственный, найденный мною, агрегатор, который общался доброжелательно, быстро отреагировал на обращение и предоставил пробный полный доступ к данным по API. На данный момент производится пробная эксплуатация и согласование взаимодействия – список организаций конкурентов и номенклатура.

Особенно удивил *4dealer.ru*. Помимо того, что данный сервис содержит программные ошибки, в чём признались разработчики ресурса, удивило отношение. Ниже скриншот обескураживающего ответа на вопрос об ошибке.



Да, и загрузить отчёт целиком за раз не получится. Предлагается использовать API для построчной загрузки (это такой мазохистский DDos).

## Добыча данных

Поскольку с *proanalytics.net* идёт утрясание моментов взаимодействия, то пришлось добывать данные с ресурса *price.ru*. Для добычи данных использовался *scraper*, спасибо *OTUS* за лекцию и домашнее задание. Но без ручного труда не обошлось. Строка формируемого датасета содержит 20 полей цен (для каждого конкурента, с сортировкой от минимальной до максимальной). При невозможности заполнить все поля, заполнение производилось минимальной ценой.

Загрузке подлежали данные о ценах конкурентов для товаров производства компании *Kyocera*. Всего было загружено данных об около 350 товарных позиция следующих категорий: *принтеры и МФУ*, *расходные материалы*, *опции и аксессуары* и *запасные части/ремонтные комплекты*. Поскольку данных было мало, то было принято решение дополнить датасет искусственными данными (произвести *аугментацию*). Были сгенерированы данные, имитирующие лежалый товар с разными периодами неактивности. Цены для того рода данных формировались на основе рекомендаций экспертов (определяли размер скидки в зависимости от периода). После выполнения аугментации размер датасета составил 912 строк.

Целевые цены были сформированы экспертами компании, без оглядки на сформированный датасет.

Дополнительно, экспертами был проанализирован датасет на предмет аномалий или некорректности данных. Часть данных была помечено как сомнительные. Забегая вперёд скажу, что обучение проводилось как с сомнительными строками, так и без них.

Сводка данных:

* Исходных данных : 228
* Augmentation : 684
* Аномалий : 144
* Всего : 912

## Текущее состояние

Для построения модели использовался датасет, структура строки (описание конкретного товара) которого, представлена ниже.

* *PMin*. Минимальная цена конкурентов на момент формирования датасета;
* *P2 - P19*. Промежуточные цены. Сформированы по возрастанию (*P2 < …< P19*);
* *PMax*. Минимальная цена конкурентов на момент формирования датасета;
* *mu*. Математическое ожидание (среднее) цены продажи конкурентами;
* *sd*. Стандартное отклонение цены продажи конкурентами;
* *md*. Медиана цены продажи конкурентами;
* *lid*. Дата последнего прихода конкретного товара на склад;
* *lip*. Цена последнего прихода конкретного товара на склад;
* *lod*. Дата последней отгрузки (продажи) конкретного товара;
* *lop*. Цена последней отгрузки (продажи) конкретного товара;
* *lsd*. Количество дней без движения – сколько дней прошло с момента последней продажи;
* *eol*. Признак «снят с производства»;
* *categoryid*. Категория товарной позиции. Использовались два варианта кодирования: как есть – числовой идентификатор, и *One Hot Encoding*.

# Решение

Решаемая задача – это задача *Регрессии*. Для решения задачи использовались два инструмента: ***LightGBM*** и ***CatBoost***.

Ниже представлена таблица результатов обучения моделей. Обучение проводилось как без аномальных данных, так и с аномальными данными.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LightGBM** | | | | | |
|  | RMSE Training | RMSE Validation | Score | OneHot | Clear |
| 1 | 2438 | 3599 | 0.9967 | X | X |
| 2 | 2346 | 3567 | 0.9967 |  | X |
| 3 | 52391 | 20098 | 0.9489 | X |  |
| 4 | 52105 | 21247 | 0.9429 |  |  |
| **CatBoost** | | | | | |
|  | RMSE Training | RMSE Validation | Score | OneHot | Clear |
| 1 | 590 | 2411 | 0.9967 | X | X |
| 2 | 584 | 3006 | 0.9967 |  | X |
| 3 | 1136 | 8143 | 0.9489 | X |  |
| 4 | 1251 | 6988 | 0.9429 |  |  |

# Дальнейшее развитие

Если получится договориться с *proanalytics.net*, а скорее всего получится, договорится о предоставлении интересующих данных, то видится, что целесообразно будет использовать нейронную сеть следующей конструкции:

Дополнительные параметры

**Подсеть дня 0  
(сегодня)**

Цены 10 конкурентов

РРЦ

**Подсеть дня -4**

Цены 10 конкурентов

РРЦ

. . .

Объединяющая подсеть

Выход