Bill Expert

Эксперт по анализу законопроектов

Шестакова Е.Д.

1 курс НИУ ВШЭ

2018

Федеральные законы (ФЗ) в нашей стране принимаются Государственной Думой (ГД), утверждаются Советом Федерации (СФ) и подписываются Президентом. Основная часть законодательного процесса происходит ГД. Существует процедура принятия законов и определенные субъекты законодательной инициативы - те, кто имеют право вносить законопроекты на рассмотрение ГД. Но далеко не все внесенные законопроекты становятся ФЗ. Так, из 1247 законопроектов, предложенных к рассмотрению Думе последнего созыва, т.е. с 5 октября 2016 года, принято всего 453.

Между тем, в обществе существует запрос на четкие "правила игры". Бизнес-сообщество хочет знать, что будет меняться в правовом поле, чтобы планировать свою деятельность. Поэтому хотелось бы иметь инструмент, предсказывающий принятие законов уже в момент внесения их в качестве законопроектов в ГД.

Особую востребованность, кажется, он будет иметь у журналистов-аналитиков. Всем нам приходилось видеть такие броские заголовки как «При замене водительских прав нужно будет повторно сдавать экзамен» или «Скоро в соцсетях придется регистрироваться по паспорту». И речь действительно шла о законодательных инициативах, правда, не прошедших в ГД. Кажется, что простая проверка моим сервисом позволила бы авторам статей формулировать заголовки более продуманно, например, «Очередная странная законодательная инициатива депутата N не имеет шансов пройти в ГД».

Таким образом, целью моего проекта является предсказание по данным о законопроекте вероятности его принятия ГД. Однако я решила анализировать не тексты законопроектов, а их названия, так как они содержат информацию об основной идее текста, но при этом значительно короче и проще для рассмотрения. Я использовала список досье на законопроекты, предоставленные компанией «Гарант». Их было около 16,5 тысяч, но важно отметить, что из них примерно 11 тысяч были отклонены и всего 5,5 тысяч приняты. Для более правильного определения результата я использовала поровну, по 5,5 тысяч, и принятых, и отклонённых законопроектов. То есть датасет составил примерно 11 тысяч названий досье.

Они имели следующий вид:

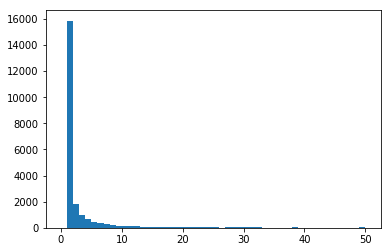
*Досье на проект Федерального закона «О внесении изменений в Федеральный закон «О контрактной системе в сфере закупок товаров, работ, услуг для обеспечения государственных и муниципальных нужд» (подготовлен Минэкономразвития России 31.12.2015)*

Также для каждого досье было указано, был ли принят соответствующий законопроект. Это бинарное значение стало целевой переменной – тем, предсказание чего является задачей проекта.

Далее я перешла к обработке данных о законопроектах. Для начала я выделила из имеющихся текстов значимые параметры: субъект законодательной инициативы (далее – автор законопроекта) и название. Далее мне нужно было проанализировать эти два вида данных по отдельности.

Различных авторов законопроектов существует всего 8, поэтому данный признак является категориальным: законопроект относится к одной из 8 категорий.

Для анализа названий, являющихся текстовыми признаками, нужно провести их предобработку. Сначала я провела токенизацию – присвоение каждому слову уникального идентификационного номера. Далее я построила диаграмму частотного распределения токенов в названиях.



На горизонтальной оси отложена частота токенов, а на вертикальной – количество токенов с данной частотой. Слова, встречающиеся в тексте редко, не могут быть использованы как параметр для определения целевого значения. Поэтому все токены с частотой меньше 10 отбрасываются, то есть заменяются нулевым токеном. Все названия имеют разные длины, но, проанализировав их распределение, можно понять, что оптимальная длина для названия составляет примерно 300 слов. Приведем все названия к данной длине: те, в которых слишком много слов, обрежем, а те, в которых недостаточно – дозаполним нулями.

Итак, после обработки данных у нас имеется около 11 тысяч законопроектов, для каждого из которых известны название, автор и бинарная целевая переменная.

Перейдем к решению задачи.

Я решила воспользоваться таким видом обучения с учителем как классифицирующая модель, предсказывающая вероятность результата. Путем проб и ошибок мне удалось подобрать наиболее удачную - Gradient Boosting. Она обучается путем формирования ансамбля более слабых моделей и оптимизации произвольной дифференцируемой функции потерь.

Gradient Boosting дает Accuracy примерно 85,5%, что я считаю довольно неплохим результатом для такого несложного метода как классифицирующая модель.

В качестве пользовательского интерфейса для своего проекта я выбрала Telegram-бота, потому что они просты как в разработке, так и в использовании, а также доступны пользователям в любое время и не требуют установки дополнительных приложений. Протестировать работу проекта можно, найдя в Telegram бота @BillExpertBot. Инструкция по использованию включена в интерфейс.

Следующим шагом в развитии моего проекта должна стать нейронная сеть, которая будет обрабатывать тексты названий при помощи алгоритма свертки. Это должно значительно повысить точность работы программы и частично преодолеть главный из её нынешних недостатков – зависимость результата от порядка слов в названии. Далее я планирую попробовать анализировать не только названия, но и полные тексты законопроектов. Однако неизвестно, как данный шаг повлияет на точность работы программы, потому что тексты содержат большое количество информации, не влияющей на результат, которая может помешать выделению значимых параметров.

Проект выполнен на Microsoft Azure Notebooks на языке Python с использованием библиотеки sklearn для машинного обучения. Также использован мессенджер Telegram для создания интерфейса. Данные для анализа предоставлены компанией «Гарант».

Всю информацию о проекте можно найти по ссылке:

github.com/oskolok-mirosdanija/BillExpert

Работа выполнена Шестаковой Екатериной, студенткой 1 курса образовательной программы «Прикладная математика» МИЭМ НИУ ВШЭ в 2018 году.