Loginom

Командное задание

Loginom Хакатон 2019

Вариант № 1

Входные данные

Представлены в файле **clients_mfo.txt** в формате **CSV**.

- Клиент – уникальный идентификатор клиента;
- Место работы – текстовое поле;

Дополнительное поле:

- Метка региона – текстовое.
- Событие (0 – не наблюдалось; 1 – наблюдалось, пусто – нет данных).

Данные от 2011-2012 года.

Постановка задачи

О клиенте микрофинансовой организации известно место его работы, которое он указывает при подаче заявки. В заявке теоретически может содержаться много полезной информации для оценки риска просрочки или невозврата займа в течение длительного времени. Но представлена данная информация в слабо структурированном виде. Хотя, очевидно, имеет смысл проверить гипотезы о связи риска просрочек с формой собственности компании, государственная она или нет, отдельно анализировать профессии с нестабильным заработком (такси, ЧОПы и т.д.). Данную информацию необходимо автоматически извлекать из названий места работы. Кроме названия места работы, известен еще город – областной центр (считаем, что город всегда известен). Это может помочь при парсинге, однако, алгоритм должен уметь перенастраиваться на другие города.

Результат

- Опубликованный компонент, принимающий на вход поле **Место работы**, дополнительные порты по усмотрению (какие-либо наборы параметров, справочники, переменные). На выходе набор из N полей (допустимо и N =1, но мы не ограничиваем этим) – переменные, производные от поля **Место работы** и имеющее существенно меньше (от 1000 до 100 раз) уникальных значений, чем исходное поле **Место работы**.
- Краткое пояснение по использованию компонента, в частности, список ограничений и требований к данным, если таковые имеются.

Оценка результата

Компонент, реализующий такую функциональность, ценен сам по себе, но дополнительно ценность полученной информации будет замеряться нами на приросте индекса AUC модели логистической регрессии, оценивающей вероятность возникновения существенных просрочек у клиентов, берущих первые и повторные займы. Поэтому мы даем поле **Событие**, которое можно использовать для оценки предсказательной силы переменных, полученных из места работы клиента средствами WoE-анализа (компонент **Конечные классы** в Loginom).

Важно

Входные данные, прилагаемые к заданию, могут рассматриваться только как частный случай. Спроектированный компонент должен работать на любом наборе данных, удовлетворяющему требуемой структуре и/или возможным ограничениям по содержащимся в нем данным (должны быть обязательно озвучены вами при их наличии). Выполнение этого требования будет проверяться нами на наборе данных из третьего региона, данных по которому мы в задании не предоставляем.

Также важно компонент сделать максимально настраиваемым, чтобы при желании пользователь смог при помощи переменных или входных справочников «тюнинговать» логику, заложенную внутри компонента. Пример справочника – аббревиатуры в названии места работы, однозначно говорящие о том, что организация государственная или бюджетная, муниципальная, частная и т.д.