ПОДТВЕРЖДЕНИЕ МАРКЕТИНГОВОЙ ГИПОТЕЗЫ С ПОМОЩЬЮ КЛАСТЕРНОГО АНАЛИЗА И МУЛЬТИКЛАССОВОЙ КЛАССИФИКАЦИИ

Панина В.В.
Томский политехнический университет, ИШИТР, 8ИМ21, e-mail: vvp39@tpu.ru
Научный руководитель: Е. А. Мирошниченко

Аннотация

Статья описывает получение, обработку и кластеризацию данных о пользователях сервиса Standuply. В статье представлено обучение мульти-классового классификатора. Результаты исследования могут быть полезны для подтверждения или опровержения маркетинговой гипотезы.

Ключевые слова: кластеризация, классификация, машинное обучение.

Введение

Сервис Standuply предназначен для автоматизации Agile-процессов компаний. В экосистеме продукта существует множество микросервисов, каждый из которых выполняет свою уникальную функцию: TODO-списки, интеграции с мессенджерами Slack и Microsoft Teams, функции ежедневного репортинга (Daily Report), а также многих других репортов из Agile-методологии (Planning Poker, Retrospective и т.д.) и т.д.. Управленческий состав компании стремится расширить воронку аквизиции для более широкого распространения продукта, получения новых пользователей и увеличения прибыли.

Отделом менеджеров была предложена маркетинговая гипотеза - «Разработка дополнительного приложения для интеграции в среду Atlassian Jira с функциональностью TODO-списков привлечет новых клиентов из сегмента крупного бизнеса, что приведет к увеличению выручки от продаж продукта».

В рамках исследовательской работы поставлены задачи анализа уже существующих пользователей и обучения модели с помощью методов машинного обучения для подтверждения или опровержения поставленной гипотезы. Разработка приложения для интеграции в Atlassian Jira уже завершена на этапе написания исследовательской работы и ведется сбор новых данных.

Подготовка данных

Данные об уже имеющихся пользователях хранятся в NoSQL базе данных MongoDB. NoSQL (Not Only SQL) – это широкий класс баз данных, который отличается от традиционных реляционных баз данных (SQL) тем, что не использует SQL для запросов и не использует табличную схему для хранения данных.

Пользователи нового приложения для интеграции – это пользователи, объединенные в группы - команды, состоящие в одном рабочем окружении в мессенджере Slack. У каждой команды может быть подписка на сервис; ограниченное число менеджеров, у которых есть доступ к управлению процессами компании через сервис личного кабинета; неограниченное число членов команды.

Количество менеджеров зависит от типа подписки и не зависит от количества членов команды.

На момент проведения исследовательской работы в базе данных содержится информация о более чем 30 000 пользователях, использующих функциональность TODO-списков. Были найдена информация о всех активных командах, в которых состоят эти пользователи. Данные собраны с помощью запросов к коллекциям базы данных и сформированы в .csv таблице, в которой содержится 7155 записей. Фрагмент таблицы представлен ниже.

Таблица 1

Фрагмент данных

Subscription	Managers_count	Team_size
free	3	11
free	3	106
free	3	24
standard	18	94
professional	3	48

Как видно, значения Subscription являются категориальными данными. Эти значения преобразованы в числовые:

Таблица 2

Фрагмент данных без категориальных данных

Subscription	Managers_count	Team_size
0	3	11
0	3	106
0	3	24
1	18	94
2	3	48

Также проведено масштабирование данных для нормализации диапазонов признаков по формуле

$$Z = \frac{(x-\mu)}{\sigma},$$

где

- Z - значение после масштабирования.
- x - исходное значение признака.
- μ - среднее значение признака (среднее значение выборки).
- σ - стандартное отклонение признака (рассчитывается автоматически с помощью библиотек языков программирования)

Получена итоговая таблица признаков.

Таблица 3

Фрагмент подготовленных данных

Subscription	Managers_count	Team_size
0	-0.130192	-0.190679
0	-0.130192	-0.141274
0	-0.130192	-0.183918
1	3.442524	-0.181318
2	-0.130192	-0.171437

Далее с помощью метода кластерного анализа K-Means (K-средних) была проведена кластеризации полученных данных для получения дополнительной информации о командах.

Количество кластеров считаем равным 3 (по гипотезе существуют 3 кластера: малый бизнес, средний бизнес и крупный бизнес).

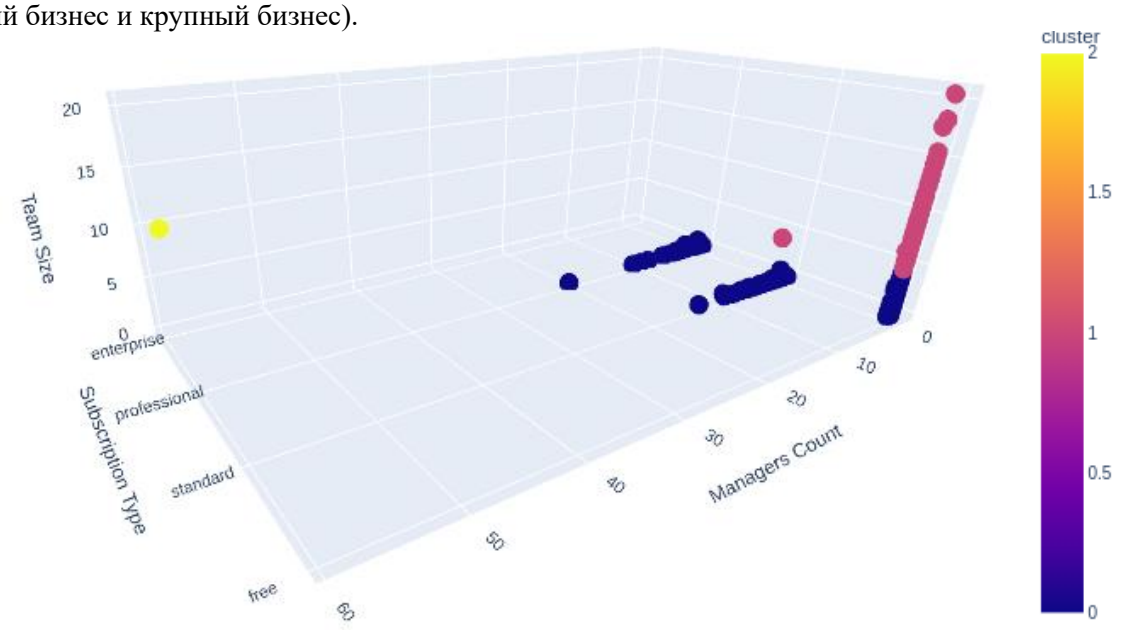

Рис. 1. Трехмерное представление данных (по кластерам)

Видно, что существует всего 1 команда из категории крупного бизнеса. Это команда компании EBay-Tech, в которой состоит 18 815 пользователей, тип подписки - Enterprise (самая дорогостоящая), количество менеджеров - 250.

Также визуализируем данные по типу подписки.

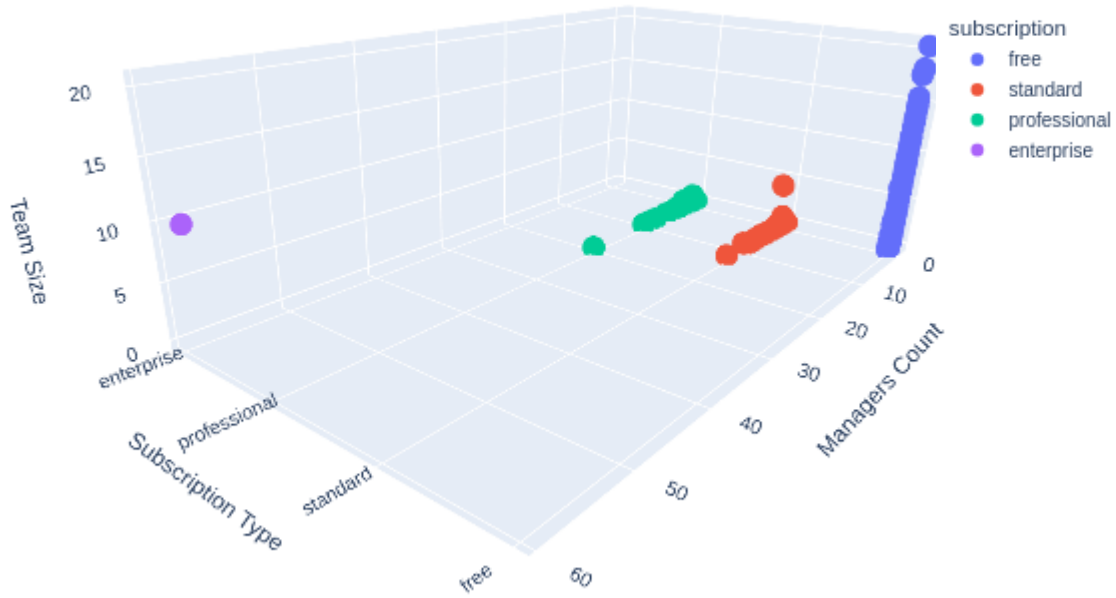

Рис. 2. Трехмерное представление данных (по типу подписки)

Видно, что продуктом пользуется большое количество команд в подпиской free (синий столбец). Количество членов таких команд может превышать 30 000, но эти команды по каким-либо причинам не используют продвинутую подписку. В таких командах до 5 менеджеров (для подписки free), которые могут управлять приложением с помощью личного кабинета. Разработка нового

приложения нацелена на привлечение клиентов с подписками professional и enterprise (зелёный и фиолетовый цвет точек на графике).

Далее полученные данные были разделены на обучающую и тестовую выборки (соотношение 70:30) для обучения мульти-классового классификатора, который будет применён в дальнейшем после развертывания приложения для Atlassian Jira и сбора данных о новых пользователях.

Обучение и результаты

Для обучения классификатора выбраны 3 модели классификаторов для того, чтобы оценить точности каждой модели и принять решение, с помощью какой модели классификатор обучать мульти-классификатор:

1. Logistic Regression – **это логистическая регрессия, которая используется** для бинарной классификации (когда нужно предсказать два класса) и многоклассовой классификации (когда нужно предсказать несколько классов, используя softmax функцию).
2. SVC – это метод опорных векторов (Support Vector Classifier), который может быть использован для бинарной и многоклассовой классификации. Он строит гиперплоскость в пространстве признаков, которая разделяет классы таким образом, чтобы расстояние от гиперплоскости до ближайших точек каждого класса было максимальным.
3. Random Forest Classifier – это случайный лес, который является ансамблевым методом машинного обучения, состоящим из нескольких деревьев решений. Он обучает каждое дерево на случайной подвыборке данных и затем принимает решение путем усреднения результатов всех деревьев.

Точность (Accuracy) – это метрика, которая показывает долю правильных прогнозов модели относительно общего числа примеров в тестовом наборе данных.

Она рассчитывается как:

$$Accuracy = \frac{\text{Количествоправильныхпрогнозов}}{\text{Общееколичествопримеровизтестовогонабора}}$$

Получены следующие значения для 3 моделей классификаторов:

Accuracy (Logistic Regression): 0.9993011879804332 или 99,93 %

Accuracy (SVM): 0.9993011879804332 или 99,93 %

Accuracy (Random Forest): 0.9986023759608665 или 99,86 %

По полученным данным можно сделать вывод, что модель работает очень точно. Для дальнейших шагов был использован метод логистической регрессии. Ниже представлены метрики, используемые для оценки производительности модели классификации.

1. Accuracy (Точность): это пропорция правильно классифицированных примеров ко всем примерам в тестовом наборе данных. Точность равна 0.9993, что означает, что примерно 99.93 % всех примеров были классифицированы правильно
2. Confusion Matrix (Матрица ошибок): это таблица, которая показывает количество верно и неверно классифицированных примеров для каждого класса (3 класс не отображен из-за недостатка данных, далее подробнее)

Таблица 4

Матрциа ошибок

	Истинный класс		
Предсказанный класс		1 класс	2 класс
	1 класс	2131	0
	2 класс	0	16

3. Precision (Точность): это пропорция правильно классифицированных положительных примеров ко всем примерам, классифицированным как положительные. Точность равна 0.9996, что означает, что примерно 99.96 % примеров, классифицированных как положительные, были на самом деле положительными.

4. Recall (Полнота): это пропорция правильно классифицированных положительных примеров ко всем истинным положительным примерам. Полнота равна 0.9583, что означает, что модель уловила около 95.83 % всех истинных положительных примеров.

5. F1 Score (F-мера): это гармоническое среднее между точностью и полнотой. Он учитывает и точность, и полноту, и вычисляется как средневзвешенное значение между ними. F1-мера равна 0.9781.

Важное замечание, что, несмотря на высокую точность модели, данных о 3 кластере (крупный бизнес) практически нет, в наборе данных всего 1 запись, которая соответствует этому значению. Также невозможно произвести кросс-валидацию из-за банального отсутствия требуемых данных. Модель высокоточно определяет данные из 1 и 2 кластеров, но для дальнейшего анализа новых пользователей этот подход не верен, ведь мы ожидаем увидеть пользователей из крупного бизнеса. В итоге такая модель подходит для опровержения гипотезы, а не подтверждения.

Как альтернативный вариант, можно рассмотреть обучение не мультиклассового классификатора, а бинарного (на основе данных из 1 и 2 кластера). В таком случае исключим единственное значение 3 кластера, что приведёт к потери информации и, возможно, к неадекватному поведению модели.

Вторая альтернатива - провести аугментацию данных 3 кластера (крупного бизнеса). Конечно, дополненные данные могут быть менее надежными или неоднородными по сравнению с исходными данными, что может привести к переобучению или потере обобщающей способности модели, но при этом будет сохраняться вся информация набора данных.

Заключение

Таким образом, проведена кластеризация и классификация пользователей, использующих функциональность TODO-листов. Классификатор будет улучшаться и использоваться в дальнейшем после сбора данных новых пользователей в интегрированном приложении для подтверждения или опровержения гипотезы.

Сейчас высокая точность модели позволяет быть уверенными в том, что данные из 1 и 2 кластера будут, верно, классифицированы, но такой уверенности нет для 3 кластера. Поэтому, на данном этапе исследования не удастся опровергнуть или подтвердить маркетинговую теорию, которая гласит, что «разработка дополнительного приложения для интеграции в среду Atlassian Jira с функциональностью TODO-списков привлечет новых клиентов из сегмента крупного бизнеса, что приведет к увеличению выручки от продаж продукта».

Исследование продолжится в рамках выпускной квалификационной работы магистра.

Список использованных источников

1. Документация Python [Электронный ресурс]. – Режим доступа: https://www.python.org/doc/ (дата доступа: 28.03.2022).

2. Кластеризация: метод k-средних [Электронный ресурс]. – Режим доступа: http://statistica.ru/theory/klasterizatsiya-metod-ksrednikh/ – Дата доступа: 28.03.2022.
3. Кластеризация [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Кластеризация – Дата доступа: 28.03.2022.
4. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation // Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science. – 2006. – Vol. 4304. – P. 1015-1021.
5. Tharwat A. Classification assessment methods // Applied Computing and Informatics. – 2021. – Vol. 17. – No. 1. – P. 168-192.