Деревья принятия решений в задаче прогнозирования оттока клиентов.

М. С. Грошев, А. В. Аргучинцев

Цель работы построить дерево принятия решений, задача которого заключается в заблаговременном нахождении сегмента пользователей, склонных через некоторый промежуток времени отказаться от использования некоторого продукта или услуги. Точное и своевременное нахождение таких пользователей позволяет эффективно бороться с их оттоком, например, выявлять причины оттока и принимать меры по удержанию клиентов.

Эта задача актуальна для большинства организаций, оказывающих услуги в сегменте B2C и вдвойне актуальна в областях, где распространение услуги близко к отметке 100%. Хороший пример такой области – рынок мобильной связи, где насыщение уже фактически произошло, и как следствие постепенно снижается прирост клиентской базы.

Этапы решения:

1) Первичный анализ данных
2) Построение дерева принятия решений
3) Оценка качества предсказаний построенного дерева.

Для решения поставленной задачи были использованы: язык программирования Python и виртуальная машина Google Colab.

В рамках первого этапа для получения статистической и визуальной оценки данных использовались: функция describe из пакета Pandas [1], библиотека matplotlib [2], библиотека *seaborn* [3]. Было выявлено, что данные имеют 16 категориальных признаков и 3 вещественных признака. Категориальные признаки были закодированы. Анализ гистограмм показал, что целевая переменная не сбалансирована. Для визуализации целевой переменной на всех признаках был использован метод понижения размерности t-SNE библиотеки sklearn [4]. Для дальнейшей работы данные были разбиты на обучающую выборку (75%) и тестовую (25%).

В рамках второго этапа на обучающей выборке с помощью библиотеки sklearn было построено решающее дерево глубины 3. Для построения решающего дерева использовался алгоритм Cart.

В рамках третьего этапа была получена оценка качества предсказаний построенного дерева. Оценка качества предсказаний будет проводится на тестовой выборке. Так как классы не сбалансированы метрика качества accuracy бесполезна. В качестве метрики для оценки, была выбрана метрика AUC-ROC, так как устойчива к несбалансированным классам. Для оценки качества предсказаний была построена матрица ошибок и подсчитана метрика AUC-ROC. Точность прогноза составила 81.5% на тестовой выборке.

Для улучшения качества прогноза использовался ансамбль деревьев решений - градиентный бустинг.

Градиентный бустинг — это подход к построению композиций, в рамках которого:

- Базовые алгоритмы строятся последовательно, один за другим.
- Каждый следующий алгоритм строится таким образом, чтобы исправлять ошибки уже построенной композиции.

В градиентном бустинге использовались 50 деревьев глубиной 3. Точность прогноза составила 85.5% на тестовой выборке.

Литература

1. Pandas // URL: https://pandas.pydata.org/ (дата обращения: 13.04.2022).
2. Matplotlib // URL: https://matplotlib.org/ (дата обращения: 13.04.2022).
3. seaborn // URL: https://seaborn.pydata.org/ (дата обращения: 13.04.2022).
4. sklearn // URL: https://scikit-learn.org/stable/ (дата обращения: 13.04.2022).