УДК 004.942

Сенин Леонид Сергеевич,
магистрант 2 курса
специальности «Информационные системы и технологии»
Института кибербезопасности и цифровых технологий
РТУ МИРЭА

Нурматова Елена Вячеславовна,
доцент кафедры аппаратного, программного и
математического обеспечения вычислительных систем,
кандидат технических наук
Института кибербезопасности и цифровых технологий
РТУ МИРЭА
nurmatova@mirea.ru

ОСОБЕННОСТИ РЕАЛИЗАЦИИ АЛГОРИТМА ПРОГНОЗИРОВАНИЯ ПО ВРЕМЕННОМУ РЯДУ

Аннотация. *В данной статье рассказывается о сравнение моделей прогнозирования временных рядов и выбор наилучшей для дальнейшего исследования.*

Ключевые слова: *временные ряды, машинное обучение, методы прогнозирования, анализ временных рядов.*

Временной ряд – последовательность наблюдений, упорядоченная по времени: (y1, y2, ..., yn), где yt – измерения некоторой переменной в n равностоящих моментов времени t = 1, 2, ..., n. Примерами временных рядов являются регулярно регистрируемые данные о всевозможных банковских операциях, заказов в ресторанах или кафе, получение электронных писем и многое другое, которые происходят (каждый час, ежедневно, еженедельно и т.п.) Все эти процессы меняются во времени и подвержены случайным колебаниям.

Градация методов прогнозирования – это достаточно неоднозначная тема и какого-то стандарта здесь нет, поэтому можно не бояться сделать ошибку. Но первоначально хотелось бы обозначить главную особенность задачи прогнозирования временных рядов по сравнению с остальными задачами машинного обучения – предсказание целевой переменной часто основывается ни на других признаках, а на значениях самой себя за предыдущий временной период. Таким образом, для прогнозирования временных рядов применяются почти все те же модели, что и для любого другого прогнозирования, но с некоторыми, модернизациями, предназначенными учесть эту их специфику.

Странно полагать, что временные ряды в какой-то существенной мере зависят от своих значений в предыдущем периоде. Большую роль будут играть внешние факторы. Именно поэтому тяжело или даже невозможно предсказать

цены акций на фондовом рынке, так как во многом она зависит от внешних факторов, такие как человек, если мы обойдёмся одними лишь временными рядами. Передовые модели должны опираться на материалы внешних факторов, такие как новостные сайты, мнение людей, работающих в этой области, просмотр других акций и другие факторы. Для этого в этой работе будет уделено особое значение предобработке данных поиск выбросов, аномалий и др.

Для моделирования ухода клиентов применяются всевозможные математические модели, среди них – логистическая регрессия, деревья решений, метод ближайших соседей (k-nearest neighbors algorithm, k-NN), опорных векторов (Support Vector Machines -SVM), случайного леса (random forest) более детально с методами и их преимуществами и недостатками можно ознакомиться ниже (табл. 1).

Таблица 1. Модели и методы для прогнозирования

Модели и методы	**Преимущества**	**Недостатки**
Бинарная регрессия	Проста. Быстро получаема. Хорошо интерпретируема. Широко применима. Достаточно точна. Обладает инструментами оценки качества моделей.	Имеет трудности из-за нелинейности отношений между оттоком и влияющими на него факторами. Предсказываемый параметр, как правило, число из непрерывного диапазона.
Метод k-ближайших соседей	Простота. Хорошо интерпретируем.	Высокая сложность одного прогноза. Проклятие размерности.
Модель выживаемости	Способны работать с цензурированными данными и категориальными переменными. Возможность визуализации	Цензурированность данных уменьшает выборку, вследствие чего могут дать несостоятельные результаты.
Нейронные сети	Устойчивы к шумам. Решают задачи при неизвестных закономерностях. Переучиваются при изменении среды. Быстродейственны. Отказоустойчивы.	Возможная неясность причин принятого решения. Отсутствие гарантии получения однозначных повторяемых результатов.
Поиск ассоциативных правил	Находит простые и интуитивно понятные правила	Выявление часто встречающихся наборов элементов требует больших вычислительных и временных ресурсов
Случайный лес	Эффективен при работе с данными с большим числом признаков и классов. Работает с непрерывными и дискретными признаками. Нечувствителен к монотонным преобразованиям.	Склонен к переобучению в случае зашумленности данных. Большая размерность моделей.

После подробного рассмотрения всех методов, которые имеют свои особенности и ограничения, а также преимущества. Можно отметить, что нейронная сеть имеет больший потенциал к рассмотрению, так как она самообучаема и устойчива к зашумленным данным, но как показывает практика из моей прошлой работы, результаты не всегда могут быть эффективными для прогнозирования оттока клиентов. Если рассмотреть метод случайных лесов решений, то они более эффективны в задачах классификации и кластеризации, что важно при моделировании оттока клиентов, однако модели имеют большую размерность данных.

На основании вышесказанного проведенного анализа, можно сказать, что наиболее широкие возможности для решения подобных задач, а именно оттока клиентов имеют такие модели как выживаемости и бинарная регрессия, позволяющие определить факторы, воздействующие на отток клиентов, и рассчитать риск наступления отказа клиента от услуг. Так же я бы рассмотрел симбиоз двух моделей, например модель случайных лесов, чтобы выделить подгруппы и бинарную регрессию.

На рис. 1 и 2 показаны экспериментально использованные стратегии, а также схема этапов обработки данных для выполнения анализа временного ряда.

В работе использован язык программирования Python, так как он является свободным и имеет большое количество библиотек, позволяющих строить, обучать и прогнозировать модели машинного обучения.

Рисунок 1. Примерная модель экспериментальной стратегии анализа временных рядов

Рисунок 2. Схема этапов обработки данных для анализа временных рядов

Наиболее эффективные библиотеками для разработки алгоритмов прогнозирования являются:

1. Tensorflow: эта библиотека, разработанная Google, широко используется при написании алгоритмов машинного обучения и выполнении сложных вычислений с использованием нейронных сетей.

2. Scikit-Learn: библиотека Python, связанная с NumPy и SciPy. Он считается одной из лучших библиотек для работы со сложными данными.

3. NumPy: Эта библиотека python, специально используемая для вычисления научных / математических данных.

4. Keras: Данная библиотека упрощает реализацию нейронных сетей. Она также обладает лучшими функциональными возможностями для вычислительных моделей, оценки наборов данных, визуализации графиков и многого другого.

5. Pandas: Эта высокоуровневая библиотека позволяет строить сводные таблицы, выделять колонки, использовать фильтры по параметрам, выполнять группировку по параметрам, запускать функции (сложение, нахождение медианы, среднего, минимального, максимального значений), объединять таблицы и многое другое. В pandas можно создавать и многомерные таблицы.

6. Matplotlib: библиотека на языке программирования Python для визуализации данных двумерной и трёхмерной графикой. Получаемые изображения могут быть использованы в качестве иллюстраций в работе.

7. SciPy: Библиотека SciPy – один из его компонентов, который включает средства для обработки числовых последовательностей, лежащих в основе моделей машинного обучения: интеграции, экстраполяции, оптимизации и других.

В реализуемом проекте также потребуются наборы отрытых данных (dataset) для машинного обучения и прогнозирования. Три основных источника,

где можно найти нужные данные: Kaggle, Data World и UCI Machine Learning Repository.

Проект разрабатывается в инструментальной среде Jupyter-ноутбук – это мощный инструмент для разработки и представления проектов Data Science в интерактивном виде. Он объединяет код и вывод все в виде одного документа, содержащего текст, математические уравнения и визуализации. Такой пошаговый подход обеспечивает быстрый, последовательный процесс разработки, поскольку вывод для каждого блока показывается сразу же.

В результате выполнения всех выше поставленных задач, ожидается, что в данном проекте на базе оптимального метода машинного обучения будет спрогнозирован отток клиентов с наибольшей точностью.

Список источников и литературы

1. Джоши, Пратик. Искусственный интеллект с примерами на Python. Пер. с англ. СПб.: ООО «Диалектика», 2019. 448 с.

2. Бокс, Дж. Анализ временных рядов прогноз и управление (часть 2) / Дж. Бокс, Г. Дженкинс. М., 2016. 207 с.

3. Лутц М. Изучаем Python, том 2, 5-е издание. Пер. с англ. СПб.: ООО «Диалектика», 2020. 720 с.