Секция «Искусственный интеллект и цифровая трансформация в бизнесе и государственном управлении»

Применение гибридных моделей для прогнозирования временных рядов, описывающих динамику вакансий

Научный руководитель – Березкин Дмитрий Валерьевич

Цаплин Сергей Тимофеевич

Аспирант

Московский государственный технический университет имени Н.Э. Баумана, Информатика и системы управления, Москва, Россия

E-mail: serjio.tsa@yandex.ru

Введение. Определение перспективных направлений развития, в которые стоит инвестировать ресурсы государства, компаний и людей, является актуальной и сложной задачей[1,2]. Социально-экономическое и научно-техническое прогнозирование являются компонентами решения этой задачи. В настоящее время существуют различные информационные ресурсы, предоставляющие общедоступные данные, которые можно использовать для такого рода прогнозирования. В качестве примера можно указать сайты по поиску работы. На таких ресурсах можно отслеживать количественные и качественные показатели появления новых профессий и специальностей. Для прогнозирования также можно использовать атлас новых профессий, который выявляет тенденции появления новых направлений специальностей в различных отраслях, и различные годовые отчеты компаний и государственных структур, в которых можно отследить социально-экономические тенденции. Методы, основанные на ручном мониторинге источников информации, являются ресурсоемкими и негибкими. Основным форматом информации, на основе которой будет осуществляться дальнейший прогноз, является временной ряд. Основная задача, решаемая в рамках исследования, заключалась в тестировании и определении наиболее эффективных методов прогнозирования временных рядов, построенных для количества опубликованных на сайте вакансий.

Основная часть.

В зависимости от используемого математического аппарата для прогнозирования временных рядов выделяют три группы методов: статистические, детерминистические и гибридные. Статистические методы основаны на функциональных зависимостях, выявленных во временных рядах. Примеры таких методов — скользящее среднее[3] и модель Холта-Винтерса[4,5,6]. Детерминистические методы, такие как методы машинного обучения, включают подходы, направленные на минимизацию ошибки прогноза. Гибридные[7] методы объединяют первые два подхода.

Обсуждение подходов:

SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) - это обновленная версия модели ARIMA [8]. Это означает, что к авторегрессионной составляющей (AR) и скользящему среднему (MA) добавляются сезонность и экзогенные параметры. Модель SARIMAX имеет 7 порядков, разделенных на две части - сезонные и несезонные: SARIMAX (p, d, q) (P, D, Q, s) Несезонные - это лаги ARIMA, к которым мы уже привыкли: p, d и q. Остальные - сезонные: P, D, Q и s. Первые три, очевидно, являются сезонными эквивалентами p, d и q, a s - единственный новый. Он обозначает продолжительность сезона, отсюда и название - «s».

XGBoostRegressor:

Основная идея алгоритма заключается в создании нескольких слабых моделей, которые решают отдельные задачи, чтобы коллективно получить одну сильную модель[9,10]. Алгоритм использует деревья решений, построенные по жадному принципу, с использованием критериев для выбора признаков. В XGBoost используются улучшенные версии деревьев решений, которые поддерживают различные стратегии разбиения и регуляризации. Для задач регрессии XGBoostRegressor может использовать различные функции потерь, такие как квадратичная ошибка, что позволяет адаптировать модель под конкретные задачи.

Гибридные модели:

Существует три подхода к созданию гибридных моделей: стекинг, бэггинг и бустинг. Стекинг заключается в передаче результатов прогнозирования нескольких моделей в одну мета-модель. Основная идея заключается в объединении сильных сторон различных моделей. Бэггинг предполагает обучение нескольких однородных моделей и усреднение их результатов. Основная идея заключается в уменьшении дисперсии и повышении устойчивости модели. Бустинг заключается в интерактивном добавлении слабых моделей для улучшения базового прогноза. Основная цель — уменьшение смещения модели.

Эксперименты:

Для эксперимента было решено использовать гибридную модель на основе стекинг-подхода. В качестве базовой модели использовалась модель SARIMAX, в качестве метамодели - xgboostregressor. В качестве исходных данных использовался набор данных, содержащий данные о размещении объявлений о работе на сайте hh.ru за 2006-2020 годы [11]. Изначально набор данных состоял из строк, содержащих подробную информацию о каждой конкретной вакансии. Набор данных включает информацию о специализации каждой вакансии, требуемом опыте, дате создания вакансии и дате ее публикации, а также различную дополнительную информацию. Доступная информация представлена на рисунке 1.

В колонке с датой публикации вакансий дата указывалась с точностью до секунд. Для более информативного анализа данных было решено отказаться от временной части даты публикации. Для этого необходимо было применить функцию преобразования ко всем наборам данных.

Для оценки прогностической способности гибридных методов было решено делать предсказания для каждого года по следующей схеме:

- Считывается очередной файл данных;
- Корректируется колонка с пометкой о дате публикации вакансии;
- Данные фильтруются для выбора интересующих нас специальностей;
- Строки группируются по дате таким образом, чтобы количество вакансий, опубликованных в определенный день, соответствовало дате этого дня;
- Модель xgboostregressor обучается для прогнозирования, начиная с 20 сентября;
- Строится модель SARIMAX, выделяются AR- и MA-компоненты;
- Значения, полученные на предыдущих этапах, добавляются в качестве признаков к исходным данным;
- Строится новая модель xgboostregressor;
- Сравниваются результаты.

В качестве моделей для оценки были выбраны четыре модели. Первой была SARIMAX, затем была обучена модель xgboost, а также 2 гибридные модели: с использованием признаков, которые мы можем извлечь из даты публикации. Для прогноза были выбраны временные ряды публикаций IT-вакансий. В качестве метрик для оценки были выбраны

MAE и MAPE (чтобы учесть масштаб исходных данных при оценке ошибки).

Результаты:

Для оценки качества прогностических моделей были выбраны две метрики. MAE - одна из известных n-масштабных зависимых мер. Если мы минимизируем ее, то получаем прогнозы, соответствующие медиане. Средняя абсолютная ошибка в процентах (MAPE) - это широко используемая мера ошибки, не зависящая от масштаба. По метрике MAE в 7 случаях лучше всего показал себя XGBoostRegressor, в 4 случаях — гибридная модель XGBoost+SARIбез передачи признаков, сформированных из даты, и в 4 случаях — гибридная модель XGBoost+SARIMAX.

Выводы. Лучшим методом оказался метод машинного обучения XGBoostRegressor[6]. Тем не менее из полученных данных следует, что дополнительные признаки, полученные из модели SARIMAX, в некоторых случаях помогают добиться более точного результата, причем именно в этих примерах. В то же время в тех примерах, где гибридная модель показывает менее точный результат, разница с лучшим результатом составляет несколько процентов и может быть принята за ошибку. Полученные результаты показывают, что можно прогнозировать уровень спроса на специалистов в той или иной области, что поможет скорректировать бюджет и программы обучения в соответствии с потребностями бизнеса и государства.

Источники и литература

1) Лавриненко Я., Тинякова В., Калашников А., Новиков А. Социально-экономические особенности регионов как фундаментальный фактор их долгосрочного развития // E3S Web of Conferences. – 2022. – Т. 110(438) – С. 02138.
2) KARAAHMETOĞLU, Ebru & ERSÖZ, Süleyman & Turker, Ahmet & Ates, Volkan & Inal, Ali. (2021). Evaluation of Profession Predictions for Today and the Future with Machine Learning Methods Emperical Evidence From Turkey..// Journal of Polytechnic. №26(1). – С 107-124.
3) Mohd Razali, N.H., Abdullah, L., Ab Ghani, A.T. et al. Exponentially Weighted Moving Average Charts Based on Interval Type-2 Fuzzy Numbers: Analyses of Quality Control and Performance. // .Int. J. Fuzzy Syst. – 2024
4) Shastri S., Sharma A., Mansotra V., Sharma A., Bhadwal A., Kumari M. A Study on Exponential Smoothing Method for Forecasting // International Journal of Computer Sciences and Engineering. – 2015. – Т. 6(4). – С. 482–485.
5) Booranawong T., Booranawong A. Double exponential smoothing and Holt-Winters methods with optimal initial values and weighting factors for forecasting lime, Thai chili and lemongrass prices in Thailand // Engineering and Applied Science Research. – 2018. – Т. 45. – С. 32–38.
6) Rahman Md. H., Salma U., Hossain Md. K., Tareq Md F. Revenue Forecasting using Holt-Winters Exponential Smoothing // Research & Reviews: Journal of Statistics. – 2016. – Т. 5. – С. 19–25.
7) Sulandari W., Yudhanto Y. Forecasting trend data using a hybrid simple moving average weighted fuzzy time series model // International Conference on Science in Information Technology (ICSITech). – 2015. – С. 303–308.
8) Arunraj N. S., Ahrens D., Fernandes M. Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry // International Journal of Operations Research and Information Systems. – 2016. – Т. 7(2). – С. 1–21.

9) Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // the 22nd ACM SIGKDD International Conference. – 2016.

10) Jackson Au, Javier Saldaña Jr, B. Spanswick, Santerre J. Forecasting Power Consumption in Pennsylvania During the COVID-19 Pandemic: A SARIMAX Model with External COVID-19 and Unemployment Variables // SMU Data Science Review. – 2020. – T. 3(2).

11) Берсенев А., Созыкин А., Шадрин Д., Кошелев А., Куклин Е., Аксенов А. Вакансии в сфере ИТ с hh.ru, 2006-2020 // IEEE Dataport. – 2021. – doi: https://dx.doi.org/10.21227/6maz-wb22.

Иллюстрации

```
Index(['id', 'description', 'key_skills', 'schedule_id', 'schedule_name',
       'accept_handicapped', 'accept_kids', 'experience_id', 'experience_name',
       'specializations', 'contacts', 'billing_type_id', 'billing_type_name',
       'allow_messages', 'premium', 'driver_license_types',
       'accept_incomplete_resumes', 'employer_id', 'employer_name',
       'employer_vacancies_url', 'employer_trusted', 'employer_alternate_url',
       'employer_industries', 'response_letter_required', 'type_id',
       'type_name', 'has_test', 'response_url', 'test_required', 'salary_from',
       'salary_to', 'salary_gross', 'salary_currency', 'archived', 'name',
       'insider_interview', 'area_id', 'area_name', 'area_url', 'created_at',
       'published_at', 'address_city', 'address_street', 'address_building',
       'address_description', 'address_lat', 'address_lng', 'alternate_url',
       'apply_alternate_url', 'code', 'department_id', 'department_name',
       'employment_id', 'employment_name', 'prof_classes_found',
       'terms_found'],
      dtype='object')
```

Рис. : Рисунок 1. Перечень доступных