МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

НАУКАСТИНГ ПРЕДСКАЗАНИЕ БЕЗРАБОТИЦЫ СРЕДИ НАСЕЛЕНИЯ НА ОСНОВЕ ЗАПРОСОВ В ПОИСКОВЫХ СИСТЕМАХ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Глуховой Алины Анатольевны

Научный руководитель
доцент, к. ф.-м. н. _____ Д. В. Мельничук

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2024

ВВЕДЕНИЕ

Актуальность темы. Последние несколько лет ознаменовались для участников экономической деятельности стремительными изменениями всех тех условий, которые раньше казались незыблемыми.

Когда ситуация может кардинально измениться буквально за несколько дней, требуется не только готовность к быстрым решениям, но и доступ к оперативно обновляемым источникам данных.

Привычные методы оценки экономических показателей дорогостоящие и не всегда эффективны: пока данные собираются и обрабатываются, экономическая ситуация может кардинально измениться.

Таким образом, задача встраивания в систему мониторинга социально-экономических показателей безработицы инструментов текущей оценки и краткосрочного прогнозирования является весьма актуальной в условиях отсутствия ряда краткосрочных статистических данных и наличия временного лага при выпуске значительного числа фактических статистических данных .

Цель работы - оценка возможностей и перспектив использования современного инструментария, в частности, наукастинга, для разработки альтернативных краткосрочных показателей в целях оперативного мониторинга такого социально-экономического индикатора как безработица на основе данных запросов поисковых систем.

Для достижения поставленной цели необходимо решить следующие задачи:

- провести анализ текущего состояния занятости населения;
- рассмотреть метод наукастинга, его суть, основные понятия и инструменты;
- выполнить анализ данных, полученных из поисковой системы Google Trends;
- построить модель регрессии данных смешанной частоты для прогнозирования уровня безработицы на основе полученных данных;
- оценить полученную модель для прогнозирования уровня безработицы;
- визуализировать результаты прогнозирования с помощью библиотеки Streamlit.

Практическая значимость. Данная работа может быть использована для решения задачи оперативной оценки хода реализации государственных программ Российской Федерации и национальных проектов в условиях отсрочки в выпуске части фактических данных, требуемых для расчета социально-экономических показателей в условиях увеличивающейся волатильности, неопределенности и нарастающего санкционного давления.

Структура и содержание работ Дипломная работа состоит из введения, пяти разделов, заключения и списка используемых источников, содержащего 20 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТ

Во **введении** раскрывается актуальность темы работы, формулируется цель работы, а также задачи, которые необходимо решить и практическая значимость результатов.

В **первом** разделе проводится анализ текущего состояния занятости населения на основе данных, полученных на сайте Федеральной службы государственной статистики из раздела "Трудовые ресурсы, занятость и безработица.

Безработица – одна из главных проблем современного общества, имеющая множество негативных последствий. Проблема роста уровня безработицы в России приобретает особую актуальность в условиях кризисного состояния национальной экономики, а также введения антироссийских экономических санкций со стороны недружественных государств.

В ходе анализа были рассмотрены данные за 2020-2023 год. Было установлено, что за указанный период времени пик роста рассматриваемого показателя пришелся на 2020 год(5,8%), что связано с повышением пенсионного возраста, поступлением на рынок труда дополнительной рабочей силы, пандемией короновирусной инфекции и экономические санкции западных государств. Самый низкий показтель установлен за 2023 год(3,2%).

В настоящий момент после ударов кризисов Российская экономика выходит на сбалансированный рост. Положение безработицы все еще остается на прежнем уровне, но будет менять в сторону преодоления кадрового кризиса.

Кроме того, в разделе рассматриваются факторы, влияющие на изменения уровня безработицы:

— факторы, повлиявшие на рост уровня безработицы (2020 год);
— факторы, повлиявшие на снижение уровня безработицы (2021-2023гг.)

В результате проведенного анализа были сделаны выводы о том, что в современных условиях российский рынок труда нестабилен и подвержен нежелательным изменениям, которые формируются под влиянием политических, правовых и демографических факторов.

Во **втором** разделе приводится теоретическая основа метода наукастинг, рассматривается определение метода наукастинг и инструменты работы с ним, его применение в различных сферах и отличительные характеристики.

Наукастинг представляет собой предсказание настоящего, ближайшего будущего и недавнего прошлого состояния экономических индикаторов. В современных реалиях наукастинг может стать достойной альтернативой традиционным статистическим показателям.

Анализировать и предсказывать тенденции потребления населения при помощи методик наукастинга и простых поисковых запросов относительно просто, чуть сложнее ситуация становится, когда происходит уход от товарно-сервисной составляющей и исследователи пытаются применить методы наукастинга в рамках анализа безработицы, делового климата и иных показателей, которые зачастую не имеют прямого выражения в сетевом потоке данных.

Модели прогнозирования по методике наукастинга в первую очередь получили применение в центральных банках, которые используют оценки для мониторинга состояния экономики в режиме реального времени в качестве оперативного косвенного расчёта официальных показателей.

Рассмотрены модели подхода к наукастингу, описаны случаи их применения и отличительные характеристики:

— Авторегрессионный анализ - используется для прогнозирования в тех случаях, когда существует некоторая корреляция между значениями во временном ряду и значениями, которые предшествуют им и следуют за ними;

— Анализ ведущих индикаторов - анализ переменных, движение которых имеет прямое отношение к движению исследуемой переменной;
— Байесовская векторная авторегрессия - использует методы Байеса для оценки вектора авторегрессии модели;
— Регрессия смешанной выборки данных - модель MIDAS, также известная как регрессия смешанной выборки данных.

Помимо теоретической основы метода наукастинг в данном разделе описывается значение поисковых систем как источника данных для прогнозирования. Такой инструмент, как статистика поисковых запросов, помогает конвертировать данные в информацию, позволяя принимать обоснованные рациональные решения.

В **третьем** разделе приводится описание и анализ данных, полученных на основе статистики запросов поисковой системы Google из Google Trends.

Google Trends является публичным web-приложением корпорации Google, основанным на поиске Google, которое показывает, как часто определенный термин ищут по отношению к общему объему поисковых запросов в различных регионах мира и на различных языках.

Такой инструмент, как статистика поисковых запросов, помогает конвертировать данные в информацию, позволяя принимать обоснованные рациональные решения. Поисковый запрос представляет собой единичные слова, ключевые фразы или предложения, которые отображаются в форме конкретного сайта как конечный результат поиска пользователя. А поисковая система, в свою очередь, это общее название службы, выполняющей поиск.

В ходе работы был осуществлен поиск данных с 2004 года по текущий год по трем запросам:
— "поиск работы";
— "hh.ru";
— "найти работу";

Полученные данные являются ежемесячными, представлены в формате "csv"и состоят из двух столбцов: месяц, количество запросов. Анализ полученых данных сводится к задаче анализа временных рядов. Для этого были построены линейные графики и выявлены компоненты временных рядов, построены графики автокорреляции, определена стационарность .

Для определения компонентов временных рядов запросов поисковых систем с помощью метода seasonal_decompose() построим график, представляющий собой аддитивную модель. В результате получим 4 графика: сам график временного ряда, тренд, сезонность и остатки.

Достоверно определить стационарность временного ряда можно с помощью критерия Дики-Фуллера.

Цель критерия Дики-Фуллера состоит в проверке наличия единичного корня во временном ряде. Единичный корень указывает на нестационарность ряда, что означает наличие систематического изменения среднего значения или тренда в данных.

Данные о уровне безработицы взяты на сайте Федеральной службы государственной статистики из раздела "Трудовые ресурсы, занятость и безработица".

Данные представлены в виде таблицы, где указан уровень безработицы, разделенные по годам и по субъктам Российской федерации.

Для построения модели были выделены данные, содержащие уровень безработицы с 2004 года по 2023 год(ежегодные данные).

Проведенный анализ позволил определить тенденции изменения количества запросов на тему безработицы среди населения и интерпретировать данную динамику.

Для этого же временного ряда были выделены его компоненты, определена стационарность с помощью теста Дики-Фуллера.

В **четвертом** разделе приводится теоретическая основа построения модели MIDAS(Mixed-Data Sampling) - модель выборки смешанных данных. Это экономные спецификации, основанные на полиномах с распределенным запаздыванием, которые гибко обрабатывают данные, отобранные с разной частотой и обеспечивают прямой прогноз низкочастотной переменной.

Одна из наиболее часто используемых параметризаций известна как "Экспоненциальная задержка Алмона поскольку она тесно связана с гладкими полиномиальными функциями задержки Алмона, которые используются для уменьшения мультиколлинеарности.

Эта функция довольно гибкая и может принимать различные формы всего с несколькими параметрами. К ним обносятся уменьшающиеся, увели-

чивающиеся или горбообразные модели.

Другая возможная параметризация, также имеющая два параметра - бета-лаг, поскольку основана на бета-функции.

Прогноз временного ряда с помощью выше указанной модели может быть осуществлен без использования агрегирования высокочастотных данныых или интерполяции низкочастотных.

MIDAS модель оценивается с помощью нелинейного метода наименьших квадратов(NLS) и agk.test - LM-тест Андреу, Гизелса, Кортеллоса. Последний выполните проверку того, равны ли гиперпараметры нормализованных экспоненциальных весов с запаздыванием по Алмону нулю.

В **пятом** разделе описывается процесс построения модели для прогнозирования уровня безработицы на основе запросов поисковых систем и создания интерфейса для визуализации полученных результатов.

Реализация практической части осуществлялась в среде разработки VisualStudio на языке Python и Rstudio на языке R.

Используемые библиотеки Python:

— pandas
— sklearn
— statsmodels
— matplotlib.pyplot
— streamlit

Используемые библиотеки R:

— quantmod
— dplyr
— ggplot2
— ggpubr
— midasr
— lubridate
— readr
— readxl
— tidyverse

Рассмотрены несколько вариантов прогнозирования данных разной частотности.

Приведение данных к одной частоте путем агрегирования высокочастотных данных является неэффективным способом, поскольку приводит к недостаточному объему выборки для построения модели.

Можно привести низкачастотные данные к высокочастотным, с помощью интерполяции. Интерполяция — это способ нахождения промежуточных значений величины по имеющемуся дискретному набору известных значений.

Интерполяция использует значения некоторой функции, заданные в ряде точек, чтобы предсказать значения функции между ними.

Для реализации прогноза были выбраны модели группы MIDAS, поскольку они предназначены для работы с временными рядами смешанной частоты.

Перед построением моделей временные ряды, являющиеся нестационарными приводятся к стационарнму виду путем дифференцирования и логорифмирования для удаления тренда и сезонности в данных, что способствует лучшему прогнозированию.

Кроме того, в работе определены формулы по которым производится расчет данных показателей.

Прогнозирование на текущий момент может быть реализовано с помощью моделирования данных смешанной частоты, избегая выше упомянутого агрегирования. Одной из таких моделей является MIDAS.

Построение моделей MIDAS(Mixed-Data Sampling) на языке R с помощью пакета "midasr". Для каждого набора данных для значимых лагов устанавливаются значение весов с помощью экспоненциальной задержки Алмона и бета-функции.

Указывается экзогенная и эндогенная переменные, которым присваиваются значения временных рядов логорифмированных и дифференцированных соответственно.

После этого строится модель midasr - базовая модель с одним высокочастотным регрессором для прогнозирования на h шагов вперед.

Параметрами модели будут также являться значимые лаги объясняющего и прогнозируемого временного ряда, определенные с помощью авторегрессии.

Далее оценивается полученный прогноз и качество модели с помощью

стандартных метрик оценки моделей временного ряда. Результаты визуализируются на графике прогноза временного ряда.

Создание интерфейса реализовано с помощью фреймворка для языка программирования Python - Streamlit.

Представленное в настоящей работе веб-приложение позволяет пользователю самостоятельно осуществить поиск необходимых данных по интересующему его запросу в Google Trends, при этом реализована возможность устанавливать временные ограничения по датам.

Для полученя данных из Google Trends необходима установка пакета Python 'pytrends'. Он позволяет создать объект

```
pt = TrendReq(hl="en-US", tz=360)
```

Далее полученные данные преобразуются в DataFrame. Ниже на странице приложения выводятся данные в виде таблицы.

Автоматически в приложении отображается график, построенный по полученным данным. На графике изображено количество запросов в течении выбранного пользователем времени.

При необходимости пользователь может скачать найденные данные на свой компбютер для работы с ними. Скачивание осуществляется после нажатия на кнопку "Download Data as csv". Соответственно данные скачиваются в csv формате.

Таким образом, данное приложение позволяет осуществить быстрый поиск необходимых данных с учетом интересующего периода времени, а также провести первичный анализ по графику и увидеть прогноз.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

— Проведен анализ текущего состояния безработицы среды населения.
— Собраны и обработы данные запросов поисковых систем, связанные с определением безработицы среди населения.
— Построены модели прогнозирования уровня безработицы на языке R (midasr) в среде разработке RStudio.
— Проведена оценка построенной модели с помощью соответсвющих метрик.

— Определены области применимости построенной модели прогнозирования безработицы.
— Реализовано веб-приложение для быстрого получения данных из Google Trends.