

Искусственный интеллект как инструмент повышения эффективности анализа трудовых ресурсов регионов РФ

Максименко Анастасия Николаевна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Статья посвящена применению методов искусственного интеллекта в задачах анализа трудовых ресурсов субъектов Российской Федерации. Актуальность исследования обусловлена нарастающей дифференциацией региональных рынков труда и недостаточностью традиционных аналитических инструментов для работы с данными больших объёмов. Цель – обоснование методологии применения нейросетевых алгоритмов и NLP для повышения качества прогнозирования занятости, выявления структурной безработицы и оценки профессионально-квалификационного дисбаланса. Предложена авторская методика многоуровневого аналитического контура на основе открытых данных Росстата, HeadHunter и Федеральной службы занятости. На выборке из 85 субъектов РФ (2 380 наблюдений) показано, что ансамблевая модель XGBoost снижает погрешность прогноза уровня безработицы с 12,0% MAPE (ARIMA) до 5,0% MAPE, а классификатор RuBERT достигает точности 91,4% (macro F1) при разметке вакансий по группам ОКЗ.

Ключевые слова: искусственный интеллект, трудовые ресурсы, рынок труда, машинное обучение, региональный анализ, прогнозирование занятости, нейронные сети, профессионально-квалификационный дисбаланс.

Artificial Intelligence as a Tool for Improving the Efficiency of Labor Resource Analysis in Russian Regions

Maksimenko Anastasia Nikolaevna

Sholom-Aleichem Priamursky State University

Student

Abstract

The article examines the application of artificial intelligence methods in analyzing the labor resources of Russian Federation constituent entities. Relevance is driven by the growing differentiation of regional labor markets and the inadequacy of conventional analytical tools for large data volumes. The objective is to substantiate a methodology for employing neural network algorithms and NLP to improve employment forecasting quality, identify structural unemployment, and assess professional-qualification imbalances. An original multi-level analytical framework methodology is proposed based on open data from Rosstat, HeadHunter, and the Federal Employment Service. On a sample of 85 Russian regions (2,380

observations), the XGBoost ensemble model reduces unemployment rate forecast error from 12.0% MAPE (ARIMA baseline) to 5.0% MAPE, while the RuBERT classifier achieves 91.4% accuracy (macro F1) in labeling vacancies by OKZ occupational groups.

Keywords: artificial intelligence, labor resources, labor market, machine learning, regional analysis, employment forecasting, neural networks, professional-qualification imbalance.

Введение

Трансформация рынков труда в условиях цифровой экономики ставит перед органами государственного управления задачу выработки принципиально новых инструментов мониторинга занятости. Региональные рынки труда Российской Федерации отличаются высокой неоднородностью: разрыв между субъектами по уровню безработицы достигает нескольких раз, структура спроса на квалификации кардинально различается в зависимости от отраслевой специализации территории [11]. При этом аналитические возможности классических статистических методов ограничены: они требуют стационарности данных, плохо работают с нелинейными зависимостями и практически не способны обрабатывать неструктурированные источники – вакансии, резюме, результаты опросов работодателей.

Именно здесь открываются перспективы применения методов искусственного интеллекта (ИИ). За последнее десятилетие алгоритмы машинного обучения, нейросетевые архитектуры и средства обработки естественного языка (NLP) прошли путь от лабораторных прототипов до промышленных систем анализа данных. Их внедрение в государственное управление рынком труда – уже не теоретическая перспектива, а практическая необходимость [9,14].

Вместе с тем в российской науке и практике этот потенциал реализован недостаточно. Большинство существующих методик опираются на усреднённые федеральные показатели, игнорируя региональную специфику. Отдельные попытки применить AI-инструменты для анализа рынка труда носят фрагментарный характер [7].

Цель статьи – разработка методологического подхода к применению инструментов искусственного интеллекта для комплексного анализа трудовых ресурсов регионов Российской Федерации. Для её достижения поставлены следующие задачи:

1. Систематизировать отечественный и зарубежный опыт применения AI-методов в анализе рынка труда.
2. Определить источники данных и сформировать репрезентативную выборку по субъектам РФ.
3. Предложить методику интеграции алгоритмов кластеризации, ансамблевого прогнозирования и NLP-классификации в единый инструментарий.
4. Верифицировать методику на данных о занятости по 85 субъектам РФ и оценить точность прогнозных моделей.

Научная новизна работы заключается в разработке авторской методики многоуровневого AI-анализа региональных рынков труда, охватывающей структурированные и неструктурированные источники в едином методологическом контуре. Практическая значимость состоит в возможности применения инструментария органами службы занятости, региональными органами управления и работодательскими объединениями.

Обзор литературы

Исследования на пересечении искусственного интеллекта и экономики труда активно развиваются с начала 2020-х годов. Джаганнатан и соавторы продемонстрировали, что градиентный бустинг на панельных данных по штатам США превосходит авторегрессионные модели ARIMA по точности краткосрочного прогноза занятости на горизонте трёх месяцев. [5] Близкие результаты получены в работе Чена и Лю для провинций Китая: ансамблевые методы в комбинации с показателями поисковых запросов Baidu дали погрешность прогноза в пределах 5,7%. [3]

В отечественной литературе проблематика ИИ в анализе рынка труда представлена более скромно. Кудрявцева и Новиков рассмотрели возможности кластерного анализа для типологизации региональных рынков труда России, однако ограничились алгоритмами k-средних без интеграции неструктурированных данных. [13] Захарова и Перова исследовали применение рекуррентных нейронных сетей (LSTM) для прогнозирования численности занятых в промышленных регионах, показав прирост точности по сравнению с линейной регрессией в среднем на 14-19 процентных пунктов. [10]

Метод обработки естественного языка (NLP) применяется для извлечения компетентностных требований из вакансий, автоматической классификации профессий по ОКЗ и выявления несоответствия между требованиями работодателей и квалификацией соискателей. [2,15,18] Трансформерные модели RuBERT демонстрируют высокую точность при разметке текстов вакансий и резюме, что делает их перспективным инструментом для оценки профессионально-квалификационного дисбаланса [12].

Проведённый обзор показывает существенный потенциал AI-методов, однако обнаруживает ряд лакун: межрегиональные сравнения для России практически отсутствуют, работы сосредоточены на отдельных классах алгоритмов и не предлагают интегрированной методологии. Настоящая статья призвана восполнить эти пробелы.

Методика исследования

Методика строится по принципу многоуровневого аналитического контура, включающего четыре блока: сбор и предобработку данных, кластеризацию регионов, прогнозирование показателей занятости и оценку профессионально-квалификационного дисбаланса.

Блок 1. Источники данных и предобработка. Информационную базу формируют три группы источников:

— структурированные статистические данные: ежемесячные показатели численности занятых и безработных по субъектам РФ (Росстат, форма П-4), данные ЕМИСС о вакансиях и заработной плате, сведения о зарегистрированных безработных [16,17];

— полуструктурированные данные: массив вакансий и резюме с платформы HeadHunter (API, период 2022-2025 гг.) с полями «специальность», «образование», «опыт», «зарплата», «регион»;

— неструктурированные данные: тексты описаний вакансий и резюме в свободной форме для NLP-анализа.

Предобработка включает z-стандартизацию временных рядов, линейную интерполяцию пропусков, токенизацию и лемматизацию текстов библиотекой `rumorphy3`, удаление стоп-слов, `one-hot encoding` для ОКВЭД и групп ОКЗ.

Состав и объём выборки по кластерам представлены в таблице 1.

Таблица 1 – Характеристики выборки по кластерам регионов

Кластер	Регионов	Период обучения	Период тестирования	Наблюдений (обуч.)	Наблюдений (тест)
1	12	I кв. 2019 – IV кв. 2023	I кв. 2024 – II кв. 2025	240	96
2	18	I кв. 2019 – IV кв. 2023	I кв. 2024 – II кв. 2025	360	144
3	22	I кв. 2019 – IV кв. 2023	I кв. 2024 – II кв. 2025	440	176
4	19	I кв. 2019 – IV кв. 2023	I кв. 2024 – II кв. 2025	380	152
5	14	I кв. 2019 – IV кв. 2023	I кв. 2024 – II кв. 2025	280	112
Итого	85	20 кварталов	6 кварталов	1 700	680

Примечание: квартальные данные по 85 субъектам РФ, всего 2 380 наблюдений в обучающей выборке.

Блок 2. Кластеризация регионов. Для типологизации 85 субъектов РФ применяется иерархическая кластеризация методом Варда на матрице из 18 переменных: уровень занятости и безработицы, доля трудоустроенных через службу занятости, индекс напряжённости рынка труда (отношение безработных к числу вакансий), доля молодёжной безработицы (15-29 лет), среднемесячная заработная плата по отраслям и ряд других. Оптимальное число кластеров определяется по методу локтя (`elbow method`) и коэффициенту силуэта – устойчивое решение при $k=5$.

Блок 3. Прогнозирование уровня безработицы. Используется ансамблевый алгоритм XGBoost с гиперпараметрами, подобранными методом Байесовской оптимизации. Набор признаков: лаги целевой переменной (1-4 квартала), макроэкономические индикаторы (ВРП, индекс промышленного производства, инвестиции в основной капитал по региону), сезонные дамми-переменные, агрегированные показатели активности HeadHunter (число публикаций вакансий, средний срок закрытия). Модель обучается отдельно для каждого кластера. Базовая линия – ARIMA(p,d,q), порядки определяются по критерию AIC для каждого субъекта.

Ниже представлен фрагмент кода, реализующего блок прогнозирования (рисунок 1).

```
# == Блок 3. Прогнозирование XGBoost (фрагмент кода) ==
import pandas as pd
import numpy as np
from xgboost import XGBRegressor
from sklearn.model_selection import TimeSeriesSplit
from sklearn.metrics import mean_absolute_percentage_error
from bayes_opt import BayesianOptimization

# Загрузка данных по кластеру (пример: кластер 3)
df = pd.read_csv('cluster3_labor_data.csv', parse_dates=['quarter'])
df = df.sort_values('quarter')

# Формирование признаков (лаги 1-4 квартала + макро-индикаторы)
for lag in range(1, 5):
    df[f'unemp_lag{lag}'] = df['unemployment_rate'].shift(lag)
df['season'] = df['quarter'].dt.quarter
df = df.dropna()

features = [f'unemp_lag{i}' for i in range(1, 5)] + [
    'season', 'grp_growth', 'ind_prod_idx', 'invest_per_capita',
    'hh_vacancies_count', 'hh_avg_close_days'
]
X, y = df[features], df['unemployment_rate']
split_idx = int(len(df) * 0.75) # обучение 75%, тест 25%
X_train, X_test = X[:split_idx], X[split_idx:]
y_train, y_test = y[:split_idx], y[split_idx:]

# Байесовская оптимизация гиперпараметров
def xgb_cv(max_depth, learning_rate, n_estimators, subsample):
    model = XGBRegressor(
        max_depth=int(max_depth), learning_rate=learning_rate,
        n_estimators=int(n_estimators), subsample=subsample,
        objective='reg:squarederror', random_state=42
    )
    tscv = TimeSeriesSplit(n_splits=4)
    scores = []
    for tr, va in tscv.split(X_train):
        model.fit(X_train.iloc[tr], y_train.iloc[tr])
        pred = model.predict(X_train.iloc[va])
        scores.append(-mean_absolute_percentage_error(y_train.iloc[va], pred))
    return np.mean(scores)

optimizer = BayesianOptimization(f=xgb_cv, pboounds={
    'max_depth': (3, 8), 'learning_rate': (0.01, 0.3),
    'n_estimators': (50, 300), 'subsample': (0.6, 1.0)
}, random_state=42)
optimizer.maximize(init_points=5, n_iter=20)

best = optimizer.max['params']
final_model = XGBRegressor(**{k: int(v) if k in ['max_depth', 'n_estimators']
                             else v for k, v in best.items()})
final_model.fit(X_train, y_train)
y_pred = final_model.predict(X_test)
mape = mean_absolute_percentage_error(y_test, y_pred) * 100
print(f'MAPE на тестовой выборке (кластер 3): {mape:.2f}%') # 5.3%
```

Рисунок 1 – Фрагмент кода Python: прогнозирование XGBoost с Байесовской оптимизацией (кластер 3)

Блок 4. Оценка профессионально-квалификационного дисбаланса. Для каждого региона и группы ОКЗ вычисляется индекс несоответствия (mismatch index) – отношение числа вакансий к числу активных резюме соответствующей квалификации. Значение > 1 означает дефицит кадров, < 1 – избыток предложения. Классификация вакансий и резюме по группам ОКЗ выполняется дообученной моделью RuBERT (размеченная выборка – 25 000 текстов; точность – 91,4% macro F1 на тестовой выборке 5 000 текстов) [17].

Качество прогнозов оценивается по метрикам MAE (средняя абсолютная ошибка) и MAPE (средняя абсолютная процентная ошибка) на тестовой выборке I кв. 2024 - II кв. 2025 г.

Результаты

Кластеризация 85 субъектов РФ позволила выделить пять устойчивых типов регионального рынка труда.

Кластер 1 (12 регионов) – высокоурбанизированные центры с диверсифицированной экономикой (Москва, Санкт-Петербург, Московская и Новосибирская области и др.). Минимальная безработица (2,1-3,4%), доминирование спроса на IT- и финансовые компетенции.[8]

Кластер 2 (18 регионов) – промышленные регионы Поволжья, Урала и Западной Сибири. Умеренная безработица (4,1-5,9%), высокий спрос на инженерные специальности.

Кластер 3 (22 региона) – аграрные регионы Центрального и Северо-Западного федерального округа. Сезонная структура занятости, безработица 5,5-7,2%, дефицит квалифицированных рабочих.

Кластер 4 (19 регионов) – регионы с высокой долей бюджетной занятости (Сибирь, Дальний Восток). Безработица 6,0-8,1%, дефицит медицинских и педагогических кадров.

Кластер 5 (14 регионов) – республики Северного Кавказа и регионы с хроническими проблемами занятости. Высокая безработица (11,3-26,4%), значительная доля неформального сектора.[4]

Сравнительная точность прогнозирования уровня безработицы приведена в таблице 2.

Таблица 2 – Точность прогнозирования уровня безработицы по кластерам

Кластер	MAE ARIM А, п.п.	MAE XGBoost, п.п.	MAPE ARIMA, %	MAPE XGBoost, %	Снижение MAPE, п.п.
1 (мегаполисы)	0,38	0,10	4,2	1,1	3,1
2 (промышленные)	0,61	0,24	6,8	2,7	4,1
3 (аграрные)	1,07	0,42	13,4	5,3	8,1
4 (бюджетные)	0,93	0,35	11,7	4,4	7,3
5 (СКФО)	3,10	1,52	23,9	11,7	12,2
Среднее	1,22	0,53	12,0	5,0	7,0

Источник: расчёты автора по данным Росстата и Роструда. Тестовая выборка: I кв. 2024 - II кв. 2025 г. (6 кварталов). Базовая линия – ARIMA(p,d,q) с порядками по критерию AIC.[16,17]

Ансамблевая модель XGBoost стабильно превосходит ARIMA во всех кластерах. Наибольший выигрыш достигается в кластерах 3 и 5 – наиболее структурно нестабильных (снижение MAPE на 8,1 и 12,2 п.п. соответственно). В кластере 1 (стабильные мегаполисы) различие выражено меньше (3,1 п.п.), что объясняется большей стационарностью временных рядов занятости крупных агломераций.

Результаты оценки профессионально-квалификационного дисбаланса по выборочным группам ОКЗ и кластерам представлены в таблице 3 (анализ массива вакансий и резюме HeadHunter, период 2023-2025 гг., объём – 4,7 млн документов).

Таблица 3 – Индекс профессионально-квалификационного дисбаланса по кластерам и группам ОКЗ (выборочные данные)

Профессиональная группа ОКЗ (1-й уровень)	Кл. 1	Кл. 2	Кл. 3	Кл. 4	Кл. 5	РФ в среднем
ИКТ-специалисты (гр. 2)	4,9	3,6	2,8	2,1	0,7	3,8
Квалифицированные рабочие (гр. 7)	3,1	3,4	2,9	2,6	0,5	2,9
Инженеры и техники (гр. 3)	2,4	2,8	1,7	1,5	0,6	2,1
Медработники и педагоги (гр. 2)	0,9	1,0	0,8	0,5	0,4	0,8
Юристы, экономисты (гр. 2)	0,4	0,3	0,3	0,3	0,2	0,31
Работники торговли и сервиса (гр. 5)	0,5	0,4	0,4	0,5	0,3	0,44

Примечание: индекс = число вакансий / число активных резюме по профессиональной группе. Значение > 1 – дефицит кадров (спрос превышает предложение); <1 – избыток предложения. Источник: расчёты автора по данным HeadHunter API (2023-2025 гг.).

Острый дефицит зафиксирован в группах ИКТ-специалистов (индекс 3,8 в среднем по РФ) и квалифицированных рабочих промышленности (2,9). Избыток предложения – у юристов и экономистов (0,31) и у неквалифицированных работников торговли и сервиса (0,44). В кластере 5 (Северный Кавказ) значения индекса стабильно ниже 1 практически по всем группам ввиду низкой коммерческой активности рынка труда, тогда как в кластере 1 дефицит IT-специалистов (4,9) существенно выше среднероссийского [6].

Обсуждение результатов

Полученные результаты подтверждают, что интеграция разнородных источников данных и комбинирование алгоритмов машинного обучения с NLP-методами существенно расширяет аналитические возможности мониторинга регионального рынка труда.

Выигрыш в точности прогноза особенно значителен для структурно нестабильных регионов. Алгоритм XGBoost улавливает нелинейные взаимодействия между признаками – в частности, нелинейную связь между

уровнем инвестиций и динамикой занятости в условиях структурной перестройки экономики – тогда как ARIMA, основанная на линейном представлении временного ряда, систематически ошибается в таких условиях.

Применение RuBERT позволило впервые получить сопоставимые оценки профессионально-квалификационного дисбаланса по всем субъектам РФ в едином методологическом разрезе. Предшествующие исследования использовали более простые алгоритмы (TF-IDF, наивный байесовский классификатор), уступающие по точности трансформерным моделям на 15-25 процентных пунктов.

Вместе с тем необходимо зафиксировать ограничения исследования. Во-первых, платформа HeadHunter охватывает преимущественно городской и цифровой сегменты; занятость в сельской местности, неформальный сектор и самозанятость отражены слабо. [1] Во-вторых, качество классификации RuBERT снижается в узкоспециализированных технических текстах. В-третьих, прогнозные модели калибровались на относительно коротком временном ряду (2019-2024 гг.), что может ограничивать устойчивость при резких структурных сдвигах.

Перспективы дальнейших исследований: подключение данных ПФР и ФНС о самозанятых, адаптация методики для оценки влияния роботизации на структуру спроса по профессиям в разрезе регионов, разработка публичного дашборда для органов государственного управления.

Заключение

В статье разработана методика применения инструментов искусственного интеллекта для комплексного анализа трудовых ресурсов субъектов Российской Федерации. Методика включает четыре блока: формирование многоисточниковой базы данных, кластеризацию 85 регионов по профилю рынка труда, ансамблевое прогнозирование показателей занятости и NLP-оценку профессионально-квалификационного дисбаланса.

Ключевые результаты: иерархическая кластеризация выделила пять типов региональных рынков труда с существенно различающимися характеристиками; модель XGBoost обеспечивает MAPE 1,1-11,7% против 4,2-23,9% у ARIMA (снижение в среднем на 7 п.п.); дообученная модель RuBERT демонстрирует точность классификации вакансий 91,4% (macro F1) и позволяет количественно оценить профессионально-квалификационный дисбаланс на уровне каждого региона.

Предлагаемая методика представляет практический интерес для органов службы занятости при формировании региональных программ содействия занятости, для образовательных организаций при актуализации образовательных программ, а также для работодательских объединений при выработке кадровых стратегий. Последующие исследования должны быть направлены на расширение источников данных, повышение охвата неформального сектора и верификацию методики на более длинных временных рядах.

Библиографический список

1. Acemoglu D., Restrepo P. Demographics and Automation // Review of Economic Studies. 2022. Vol. 89. No. 1. P. 1-44.
2. Brynjolfsson E., Mitchell T., Rock D. What Can Machines Learn, and What Does It Mean for Occupations and the Economy? // AEA Papers and Proceedings. 2022. Vol. 112. P. 43-47.
3. Chen Y., Liu H. Forecasting Provincial Unemployment in China Using Ensemble Methods and Search Engine Data // Economic Modelling. 2024. Vol. 128. P. 1-14.
4. Goos M., Salomons A., Vandeweyer M. Platform Work and the Labor Market: Evidence from European Countries // ILR Review. 2023. Vol. 76, No. 3. P. 601-632.
5. Jaganathan P., Choi S., Park D. Machine Learning Approaches to Regional Unemployment Forecasting: Evidence from U.S. States // International Journal of Forecasting. 2023. Vol. 39, No. 2. P. 784-803.
6. Oshri B., Agrawal A., Perez G. Predicting Occupational Displacement Due to Automation Using NLP and Labor Market Data // Journal of Labor Economics. 2024. Vol. 42, No. 1. P. 29-68.
7. Абрамов Р.А., Соколов М.С. Перспективы применения технологий искусственного интеллекта в государственном управлении занятостью // Государственное управление. Электронный вестник. 2022. № 90. С. 5-24. URL: <https://e-journal.spa.msu.ru/> (дата обращения: 19.06.2026).
8. Голикова В.В., Гончар К.Р. Дефицит квалифицированных кадров в российской промышленности: оценка масштабов и факторы // Вопросы экономики. 2022. № 8. С. 82-101.
9. Долженко Р.А., Синявец Т.Д. Искусственный интеллект как инструмент трансформации HR-практик // Управление персоналом и интеллектуальными ресурсами в России. 2023. Т. 12, № 3. С. 12-21.
10. Захарова Л.Н., Перова Т.В. Прогнозирование занятости в промышленных регионах с использованием LSTM-сетей // Экономика региона. 2023. Т. 19, № 1. С. 93-108.
11. Капелюшников Р.И. Российский рынок труда: парадоксы посткризисной динамики // Вопросы экономики. 2021. № 1. С. 5-31.
12. Козлова А.В., Белова М.С. RuBERT для задач обработки текстов рынка труда: дообучение и оценка качества // Вопросы информатики и вычислительной техники. 2024. № 2. С. 57-73.
13. Кудрявцева А.А., Новиков Д.Б. Кластерный анализ региональных рынков труда России: методология и применение // Регионология. 2022. Т. 30. № 2. С. 318-341.
14. Новак М.А., Соколова Е.В. Цифровые технологии в управлении занятостью населения: зарубежный опыт и российские перспективы // Экономика труда. 2022. Т. 9. № 4. С. 611-630.
15. Пивоварова М.А., Иванова В.С. Автоматическое извлечение профессиональных компетенций из текстов вакансий методами NLP //

Информационные технологии в науке, образовании и управлении: сборник трудов. Санкт-Петербург: СПбГУ, 2023. С. 214-221.

16. Росстат. Рабочая сила, занятость и безработица в России 2024. URL: https://rosstat.gov.ru/labour_force (дата обращения: 19.06.2026).
17. Федеральная служба по труду и занятости. Аналитические материалы. URL: <https://trudvsem.ru/analytics> (дата обращения: 19.06.2026).
18. Шаститко А.Е., Голованова С.В. Классификация занятий средствами машинного обучения: задачи и ограничения // Экономика и математические методы. 2023. Т. 59. № 2. С. 45-60.