РОССИЯ И ЭКОНОМИКА ФОРМИРУЮЩИХСЯ РЫНКОВ

УДК 331.2
JEL J31+C38

Моделирование влияния социальных факторов на динамику серых зарплат*

Н. Н. Куницына[1], *Ю. А. Метель*[2]

[1] Северо-Кавказский федеральный университет,
Российская Федерация, 355017, Ставрополь, ул. Пушкина, 1
[2] Отделение по Ставропольскому краю Южного главного управления
Центрального банка Российской Федерации,
Российская Федерация, 355035, Ставрополь, ул. Ленина, 286

Для цитирования: Куницына, Н. Н. и Метель, Ю. А. (2024) 'Моделирование влияния социальных факторов на динамику серых зарплат', *Вестник Санкт-Петербургского университета. Экономика*, 40 (3), с. 460–482. https://doi.org/10.21638/spbu05.2024.306

Проблема отбеления заработной платы значительной доли россиян обострилась в период пандемии, постковидного восстановления и особенно на фоне геополитических шоков. Беспрецедентное расширение мер государственной поддержки, предоставляемых адресно и официально, побудило предпринимателей и граждан к выходу из тени. Несмотря на явные позитивные сдвиги, проблема зарплаты «в конвертах» не искоренила себя: в современных публикациях ученые возвращаются к ней вновь и вновь. При оценке данного социально-экономического явления наука опирается большей частью на методы регрессионного, панельного анализа, инструментальных переменных и проч. В настоящей работе влияние социальных факторов на динамику серых зарплат выявлено на основе методов градиентного бустинга, линейного моделирования, случайного леса и наивного байесовского классификатора. В качестве исходной информации использованы результаты обследований Российского мониторинга экономического положения и здоровья населения Высшей школы экономики (РМЭЗ НИУ ВШЭ), находящиеся в свободном доступе. В результате моделирования определена зависимость способа получения заработной платы (официальный, «в конверте») от ряда факторов, из которых наиболее сильное влияние оказали отраслевая принадлежность организации, ее размер, форма собственности, уровень образования сотрудника, про-

* Настоящая статья отражает личную позицию авторов. Содержание и результаты данного исследования не следует рассматривать, в том числе цитировать в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решение регулятора.

должительность его отпуска, удовлетворенность профессиональным ростом и условиями труда. Прикладное значение полученных результатов состоит в возможности выработки управляющих воздействий как на уровне государства, так и на уровне субъектов хозяйствования в направлении нивелирования выявленных причин получения серых зарплат и сокращения масштабов скрытого фонда оплаты труда.

Ключевые слова: заработная плата, серая зарплата, градиентный бустинг, линейные модели, случайный лес.

Введение

Пандемия коронавируса, а вслед за ней шоки от геополитических событий, безусловно, не могли не отразиться на динамике рынка труда. Колебаниям подверглись не только масштабы занятости, формальной и неформальной, но и способы получения заработной оплаты. Если в допандемийный период популярными являлись выплаты «в конвертах», ориентированные на мнимое сокращение налоговых платежей и социальных взносов, то на фоне расширения мер поддержки со стороны государства граждане, предприниматели и бизнес в целом все больше становятся приверженцами белых выплат. Как следует из исследования компании HeadHunter, доля россиян, получающих зарплату официально, в 2022 г. увеличилась до 71 % (для сравнения: в 2016 г. она составляла 57 %, в 2018 г. — 54 %, в 2020 г. — 66 %)[1], доля серой зарплаты сократилась до 19 %, черной — до 10 %. Серые зарплаты получают порядка 15 млн россиян[2]. По информации Росстата, скрытый фонд оплаты труда в 2022 г. составил 19,5 % от общего объема оплаты труда наемных работников (для сравнения: в 2021 г. — 20,3 %, в 2016 г. — 27 %)[3].

Согласно исследованиям Всероссийского центра изучения общественного мнения (ВЦИОМ), результаты несколько иные: 86 % россиян в 2021 г. сообщили, что получают всю зарплату официально (в 2006 г. — 71 %, в 2016 г. — 75 %); 6 % получают всю зарплату неофициально, «в конверте» (в 2016 г. — 10 %); 5 % получают часть зарплаты официально, а часть — «в конверте» (в 2016 г. — 13 %)[4].

Несмотря на явно позитивную динамику, к полному успокоению это явно не приводит: в итоге 29 % жителей страны получают неофициальный доход полностью или частично.

Серые зарплаты представляют проблему не только для государства, поскольку заведомо снижаются суммы поступлений в бюджеты и социальные фонды, но и для работников ограничивается размер отпускных, пособий по временной нетрудоспособности, по беременности и родам, по уходу за ребенком, выходного пособия при увольнении, будущих пенсий, сокращается сумма возможного кредитного лимита, ограничивается возможность получения социальных и имущественных налоговых вычетов. Ответственность за выплату зарплаты «в конверте»

[1] Соловьева, О. (2022) ‘Старение населения отбелило зарплаты’. *Независимая газета*. URL: https://www.ng.ru/economics/2022-07-13/1_8485_salaries.html (дата обращения: 04.06.2023).

[2] Там же.

[3] Виноградова, Е. (2023) ‘Власти обсудили, как сократить число получающих зарплату «в конверте»’. *РБК*. URL: https://www.rbc.ru/economics/28/04/2023/644a69299a79470b33e7efe1 (дата обращения: 04.06.2023).

[4] ВЦИОМ. (2016) «Белая» зарплата vs «черный нал» URL: https://wciom.ru/analytical-reviews/analiticheskii-obzor/-belaya-zarplata-vs-chernyj-nal- (дата обращения: 26.06.2023).

несет работодатель, что может повлечь дополнительные проверки, штрафы и даже уголовное наказание. Не следует забывать и о коэффициенте замещения, который, согласно заключению Счетной палаты России по проекту бюджета Фонда пенсионного и социального страхования (единого Соцфонда) в 2023–2025 гг., составит в 2025 г. лишь 30 %, а в 2026 г. опустится ниже этой величины[5], что значительно меньше 40 %, прописанных как минимально допустимая норма социального обеспечения граждан старших возрастов (в развитых странах — 60–70 %) в ратифицированной Российской Федерацией Конвенции Международной организации труда (МОТ) о минимальных нормах социального обеспечения (Конвенция № 102).

Острота отмеченных проблем побудила к обращению к причинам получения серой заработной платы. Социальные, демографические и экономические показатели, воздействуя на которые можно сокращать масштабы неофициальной оплаты труда, требуют дополнительного анализа, поскольку периоды пандемии, постковидного восстановления, расширения санкционной нагрузки, безусловно, повлияли на состояние рынка труда и трансформацию процессов легальных выплат.

В связи с этим целью исследования явилось определение факторов разнонаправленного воздействия на динамику серых зарплат, практическая значимость которого состоит в возможности выработки комплекса рекомендаций, прямых и косвенных мер, способных сократить масштабы неофициальных заработных плат в организациях и скрытого фонда оплаты труда в экономике в целом.

Достижение поставленной цели потребовало сформировать массив эмпирических данных, провести моделирование зависимости серых зарплат от ряда факторов, представить предложения по использованию результатов исследования государственными институтами с целью минимизации частоты случаев получения заработной платы «в конверте» на российском рынке труда.

1. Теоретические предпосылки исследования

Распространенным явлением на российском рынке труда является выплата серой зарплаты. Зачастую такая ситуация возникает, когда работодатель официально выплачивает сотруднику сумму, близкую к минимальному размеру оплаты труда (МРОТ), наряду с которой производит доплаты, не включаемые в налогооблагаемую базу по платежам в бюджет и взносам во внебюджетные фонды. Деньги, за которые работодатель не отчитывается перед Федеральной налоговой службой, Социальном фондом России, не перечисляемый им как налоговым агентом НДФЛ и называют серой оплатой за труд[6].

Разграничивая виды выплат, отметим, что белой называют официальную заработную плату, с полного размера которой удержан НДФЛ и уплачены все обязательные страховые взносы; серой считается заработная плата, которая состоит из официальной части, предполагающей отчисления налогов и обязательных платежей, и неофициальной, с которой не уплачиваются налоги и страховые взносы;

[5] РБК. (2022) *Пенсионное страхование теряет эффективность*. URL: https://www.rbc.ru/newspaper/2022/10/21/635117919a79476aef0ab54c (дата обращения: 28.07.2023).

[6] Скрябин, С. (2023) ‘Серая зарплата: какая ответственность работодателя’. *Бизнес-секреты*. Март 2023. URL: https://secrets.tinkoff.ru/bezopasnost-biznesa/shtrafy-za-seruyu-zarplatu/ (дата обращения: 05.06.2023).

черную зарплату выплачивают работникам без официального трудоустройства «в конверте» без каких-либо отчислений в бюджет.

Поскольку в цели настоящего исследования не входил глубокий категориальный анализ литературы в части разграничения белой, серой и черной зарплат, в данной работе серыми будут считаться выплаты работникам, не отражаемые в официальной отчетности.

Так, Федеральная служба государственной статистики Российской Федерации (Росстат) скрытую оплату труда и смешанные доходы подсчитывает балансовым методом. Из суммарных расходов россиян (включая прирост их финансовых активов) вычитаются формально зарегистрированные доходы. Показатель не отражает теневые доходы как таковые (в том числе зарплаты «в конвертах»), однако включает их[7].

Изучению данного социально-экономического явления посвящено большое количество публикаций. Многие ученые проводят исследования факторов, влияющих на динамику официальных и неофициальных выплат, используя различные методы и приемы. Так, в исследованиях А. В. Шаруниной (Шарунина, 2013), коллективов авторов под руководством В. Е. Гимпельсона (Гимпельсон и др., 2010; Гимпельсон, Капелюшников и Шарунина 2018; Гимпельсон и Капелюшников, 2020) определяются факторы, влияющие на заработную плату, и оцениваются МНК-регрессии, учитывающие эти факторы, в том числе с использованием динамической мультиномиальной логит-модели со случайными эффектами. В 2021 г. В. Е. Гимпельсон опубликовал результаты исследований бинарной пробит-регрессии в части влияния пандемии на динамику заработных плат (Гимпельсон, 2021). Пробит-регрессию и логистическую регрессию для оценки социально-экономических факторов, влияющих на желание сменить работу, применили М. В. Ляхнова и Г. С. Рудаев (Ляхнова и Рудаев, 2021). Анализ уровня удовлетворенности условиями труда с позиций его оплаты О. В. Вередюк (Вередюк, 2020) также провела на основе порядковой логистической регрессии. Данный подход, наряду с методами случайный лес и k-средних, использовала и Л. Р. Абзалилова (Абзалилова, 2021). Случайный лес отражен в методологических подходах к исследованию неформальной занятости, в том числе через призму оплаты за труд, Е. В. Заровой и Э. И. Дубравской (Зарова и Дубравская, 2020).

Еще одним методом, нашедшим применение для изучения характеристик рынка труда, в том числе заработной платы, стал дискриминантный анализ, результаты которого представлены в работе А. В. Мальцевой, О. В. Махныткиной, Н. Е. Шилкиной (Мальцева, Махныткина и Шилкина, 2015).

Кластеризацию вакансий по уровню заработной платы и занятости населения осуществили О. А. Хохлова, А. Н. Хохлова, А. Ц. Чойжалсанова (Хохлова, Хохлова и Чойжалсанова, 2022), Ю. Е. Гавриленко (Гавриленко, 2022).

И. А. Дембовский и А. А. Машков (Дембовский и Машков, 2019) провели оценку доли серых зарплат в регионах России для агентного моделирования процессов трудоустройства жителей.

[7] Виноградова, Е. (2023) 'Власти обсудили, как сократить число получающих зарплату «в конверте»'. *РБК*. URL: https://www.rbc.ru/economics/28/04/2023/644a69299a79470b33e7efe1 (дата обращения: 06.06.2023).

Т. Л. Журавлева, интегрировав накопленный наукой опыт, отмечала, что популярными в данной предметной области являются методы анализа панельных данных на основе модели с фиксированными эффектами, инструментальных переменных и регрессии с переключением, а также метод контрольной группы (Журавлева, 2015). Вместе с тем, по ее словам, «при множестве имеющихся методов… для построения наиболее полной и достоверной картины ситуации наилучшей стратегией является применение различных методологий и сравнение получаемых результатов. …Специфические особенности рынка труда наводят исследователей на мысль о создании новых инструментов».

В вопросе моделирования серых заработных плат наука опирается большей частью на методы регрессионного, панельного анализа, инструментальных переменных. В связи с этим предпринята попытка применения нелинейных методов (градиентного бустинга, случайного леса, байесовского классификатора) с целью дифференциации полученных результатов в комплексе и сравнения полученных результатов с одним из традиционных подходов.

2. Методология исследования

В целях построения моделей, отражающих зависимость серой заработной платы от описывающих переменных, был применен ряд методов:

1) градиентный бустинг (модель CatBoost);
2) логистическая регрессия (Logistic Regression);
3) линейный дискриминантный анализ (Linear Discriminant Analysis);
4) непараметрический метод обучения k-ближайших соседей (K-Nearest Neighbors);
5) случайный лес / ансамбль деревьев (Random Forest);
6) вероятностные классификаторы (Probabilistic Classifiers) — наивный байесовский классификатор (Naive Bayes Classifier).

1. **Метод градиентного бустинга** CatBoost (Gradient Bosting (Machine), GBM)[8] — алгоритм обучения с учителем для задач регрессии и классификации, который позволяет строить аддитивную функцию в виде суммы деревьев решений итерационно, по аналогии с методом градиентного спуска (Dorogush, Ershov and Gulin, 2017). Его основной идеей является устранение ошибок предыдущего шага. Обобщенный алгоритм работы предполагает определение набора данных $\{(x_i, y_i)\}_{i=1,\ldots,n}$; числа итераций M; выбор функции потерь $L(y, f)$ с выписанным градиентом; выбор семейства функций базовых алгоритмов $h(x, \theta)$ с процедурой их обучения; определение дополнительных гиперпараметров, например глубины дерева у дерева решений (Prokhorenkova et al., 2019).

Затем проводится инициализация GBM константным значением $\hat{f}(x) = \widehat{f_0}, \widehat{f_0} = \gamma, \gamma \in R$:

$$\hat{f}_0 = \text{argmin}\, \Sigma_{i=1}^{n} L(y_i, \gamma). \quad (1)$$

После чего для каждой итерации $t = 1, \ldots, M$ повторяется ряд действий:

1) определение псевдоостатков r_t:

[8] Натекин, А. (2017) ‘Открытый курс машинного обучения. Градиентный бустинг’. *Хабр*. Май 2017. URL: https://habr.com/ru/companies/ods/articles/327250/ (дата обращения: 11.06.2023).

$$r_{it} = -\left[\frac{\delta L\left(y_i, f\left(x_i\right)\right)}{\delta f\left(x_i\right)}\right]_{f(x)=\hat{f}(x)}, \quad i = 1, \ldots, n; \tag{2}$$

2) построение нового базового алгоритма $h_t(x)$ на псевдоостатках $\{(x_i, r_{it})\}_{i=1,\ldots,n}$;

3) поиск оптимального коэффициента ρ_t при $h_t(x)$ относительно исходной функции потерь:

$$\rho_t = \operatorname{argmin} \sum_{i=1}^{n} L\left(y_i, \hat{f}\left(x\right) + \rho \cdot h\left(x_i, \theta\right)\right); \tag{3}$$

4) сохранение $\widehat{f_t}\left(x\right) = \rho_t \cdot h_t\left(x\right)$.

После этого формируется итоговая GBM-модель $\hat{f}(x)$:

$$\hat{f}\left(x\right) = \sum_{i=0}^{M} \widehat{f_i}\left(x\right). \tag{4}$$

2. **Логистическая регрессия** (Logistic Regression) является одним из методов классификации с использованием линейного дискриминанта Фишера. Вероятность отнесения к той или иной группе (классу) рассчитывается по формуле

$$p = \frac{1}{1 + e^{-z}}, \tag{5}$$

где z — стандартное уравнение регрессии $z = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_{\text{n}} \cdot x_{\text{n}} + a$; x — значения независимых переменных; b — коэффициенты, расчет которых является задачей бинарной логистической регрессии; a — константа.

3. **Линейный дискриминантный анализ** (Linear Discriminant Analysis) позволяет предсказать принадлежность объектов к двум и более непересекающимся группам[9]. Его ядром выступает дискриминантная функция следующего вида:

$$d = b_1 x_1 + b_2 x_2 + \ldots + b_{\text{n}} x_{\text{n}} + a, \tag{6}$$

где x_1 и x_n — значения переменных; b_1 и b_n — коэффициенты, оцениваемые с помощью дискриминантного анализа.

Целью данного подхода является определение коэффициентов дискриминантной функции, позволяющих с максимальной четкостью провести разделение объектов на группы.

4. Ключевая идея **метода K-Nearest Neighbors** заключается в оценивании сходства объектов: конкретный объект $\overrightarrow{X_{N+1}}$ относится к тому классу Q_{k} ($k = 1, \ldots, K$), к которому принадлежит его ближайший сосед X_J^* из обучающей выборки. Решающее правило имеет вид[10]:

[9] Академия НАФИ. (2017) *Дискриминантный анализ.* URL: http://nafi.ru:8080/upload/spss/Lection_10.pdf (дата обращения: 01.09.2023).

[10] Pedregosa et al. (2011) *Scikit-learn: Machine Learning in Python, KNeighborsClassifier.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата обращения: 01.10.2023).

$$d\left(\overrightarrow{X_J^*}\,\overrightarrow{X_{N+1}}\right) = \min d\left(\overrightarrow{X_J^*}\,\overrightarrow{X_{N+1}}\right), \quad \text{для} \quad \forall j = 1, \ldots, N, \tag{7}$$

где d — мера близости (в данном исследовании использовалось Манхэттенское расстояние или расстояние городских кварталов).

5. **Метод случайного леса** (Random Forest) предусматривает построение модели $\boldsymbol{N}$ деревьев, в которой для каждого $n = 1, \ldots, N$ можно сгенерировать выборку X_n с помощью бутстрэпа и построить решающее дерево b_n по выборке X_n. Дерево строится до тех пор, пока в каждом листе не более $n_{\min}$ объектов или пока не достигается определенная высота дерева (Pedregosa et al., 2011).

Итоговый классификатор $a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x)$ позволяет выбрать решение «голосованием по большинству».

6. **Наивные методы Байеса** (Gaussian Naive Bayes, GNB) — набор алгоритмов контролируемого обучения, основанных на применении теоремы Байеса с «наивным» предположением об условной независимости между каждой парой характеристик при заданном значении переменной класса[11].

GNB реализует гауссовский наивный байесовский алгоритм для классификации. Предполагается, что вероятность появления признаков гауссова:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right). \tag{8}$$

Параметры σ_y, а также μ_y оцениваются с использованием максимального правдоподобия.

Важно отметить, что в задачах бинарной классификации дефолтное значение порога принятия решения (threshold) установлено на уровне 0,5. Данная граница часто не является оптимальной. Поэтому в настоящем исследовании для ее «тонкой настройки» выбрана точка пересечения (по оси X) False-Positive Rate (FPR) и False-Negative Rate (FNR). False-Positive и False-Negative — ошибки классификации. В статистике первый вид ошибок принято называть ошибкой 1-го рода, а второй — 2-го рода. В настоящем исследовании ошибкой 1-го рода будет принятие официальной заработной платы за серую, так как нулевая гипотеза состоит в том, что никто из индивидов не получает оплату за труд таким образом, и мы эту гипотезу отвергаем. Следовательно, ошибкой 2-го рода будет являться пропуск индивида, получающего серую заработную плату, и ошибочное принятие нулевой гипотезы.

Для расчета FPR и FNR использованы формулы:

$$FPR = \frac{FP}{FP + TN}, \tag{9}$$

[11] Pedregosa et al. (2011) *Scikit-learn: Machine Learning in Python, Gaussian Naive Bayes. GaussianNB.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html (дата обращения: 22.09.2023).

$$FNR = \frac{FN}{FN + TP}, \tag{10}$$

где *FP* — ошибочно позитивные ответы; *TN* — истинно негативные ответы; *FN* — ошибочно негативные ответы; *TP* — истинно позитивные ответы (рис. 1).

	Positive (0)	Negative (1)
Positive (0)	TP	FP
Negative (1)	FN	TN

Рис. 1. Матрица ошибок (Confusion Matrix) для False-Positive Rate и False-Negative Rate

Источник: Лабынцев, Е. (2017) Метрики в задачах машинного обучения, *Хабр*. URL: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 11.06.2023)

Вместе с тем важным этапом определения качества построенной модели является ее оценка без учета threshold. Для решения этой задачи целесообразно рассчитывать площадь ROC-AUC[12] под кривой ошибок[13] — линией в диапазоне от (0; 0) до (1; 1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR). При этом TPR вычисляется по формуле

$$TPR = \frac{TP}{TP + FN}. \tag{11}$$

Идея данной метрики состоит в том, что на построенном графике каждая отмеченная точка соответствует выбору определенного порога. Тогда площадь ROC-AUC показывает качество построенного алгоритма. При этом важность представляет и угол наклона кривой ошибок — она должна стремиться к максимальным значениям TPR и минимальным FPR, то есть к точке с координатами (0; 1)[14].

С целью определения степени и направления влияния используемых в модели факторов на эндогенную переменную применен алгоритм Шепли (Shapley Values), основанный на концепции коалиционных игр (Valouet et al., 2022)[15]. Для оценки важности предиктора происходит оценка предсказаний модели с ним и без него. Значение Шепли для *i*-го фактора рассчитывается для каждого набора данных на всех возможных комбинациях факторов (включая их отсутствие), затем полученные значения суммируются по модулю и получается итоговая важность *i*-го фактора (Geanderson et al., 2020)[16].

[12] Area Under Curve.

[13] Receiver Operating Characteristic Curve.

[14] Лабынцев, Е. (2017) 'Метрики в задачах машинного обучения'. *Хабр*. URL: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 12.06.2023).

[15] Lundberg, S. (2018) 'Welcome to the SHAP documentation'. *Shap*. URL: https://shap.readthedocs.io/en/latest/index.html (дата обращения: 13.06.2023).

[16] Трошенков, П. (2018) 'Как интерпретировать предсказания моделей в SHAP'. *Хабр*. URL: https://habr.com/ru/articles/428213/ (дата обращения: 18.06.2023).

Оценка вклада каждого параметра в принятие решения осуществлено на основе значения Шепли, рассчитываемого по следующей формуле

$$\phi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!} (p(S \cup \{i\} - p(S)), \quad (12)$$

где S — произвольный набор факторов без i-го фактора; n — количество факторов; $p(S \cup \{i\})$ — предсказание модели с i-м фактором; $p(S)$ — предсказание модели без i-го фактора.

3. Вычислительные эксперименты и их анализ

Согласно поставленной цели исследования, проведено моделирование факторов, популяризирующих серую заработную плату описанными методами. Случаи выплаты заработка частично «в конверте» и полностью «в конверте» объединены в одну категорию «Неофициальная (серая) заработная плата».

Проанализированы результаты опросов Российского мониторинга экономического положения и здоровья населения Научно-исследовательского института Высшей школы экономики (РМЭЗ НИУ ВШЭ[17]) за период с 1994 по 2022 г., включающего 2775 переменных и 233 732 наблюдений. При этом не все переменные РМЭЗ использованы в моделировании, а лишь те из них, которые оказывают влияние на характер оплаты за труд *type_of_salary*. Многие вопросы РМЭЗ имеют большое количество пропущенных значений, поэтому в ходе предварительной обработки первоначальной выборки из нее были удалены пропущенные значения и опечатки. В итоге количество наблюдений в разрезе 29 факторов (табл. 1) составило 13 567 ед.

Первоначально выборка была разделена на тренировочную и тестовую в пропорции 75:25 %. При этом тестовая не была задействована в обучении моделей и предназначена исключительно для оценки их качества.

Затем описывающие переменные в каждой выборке разделены на две группы: категориальные (номинативные) и непрерывные, после чего проведена их стандартизация — масштабирование путем преобразования статистического распределения в формат со средним значением, равным нулю, и стандартным отклонением, равным единице по формуле

$$z = \frac{x-u}{s}, \quad (13)$$

где z — стандартизированное значение; x — текущее значение показателя; u — среднее значение выборки; s — стандартное отклонение.

[17] Все данные обследований РМЭЗ ВШЭ, начиная с 1994 г. и до сегодняшнего дня, находятся в свободном доступе. Данные РМЭЗ ВШЭ имеют панельную структуру, однако они получены на основе опросов и, следовательно, могут преднамеренно или непреднамеренно искажаться респондентами; *HSE.* URL: http://www.hse.ru/rlms (дата обращения: 11.05.2023).

Таблица 1. **Описание переменных, используемых в моделировании**

№	Показатель	Сокр.	Вариант ответа
1	Характер оплаты за труд [18]	*type_of_salary*	0 — всю заработную плату получаю официально; 1 — часть заработной платы получаю официально, часть — нет; либо заработная плата в полном объеме выплачивается неофициально
2	Регион	*region*	указывается респондентом
3	Семейное положение	*marst*	1 — женат (замужем); 2 — не женат (не замужем)
4	Образование	*diplom*	1 — окончил 0–6 классов; 2 — незаконченное среднее образование (7–8 классов); 3 — незаконченное среднее образование (7–8 классов) и что-то еще; 4 — законченное среднее образование; 5 — законченное среднее специальное образование; 6 — законченное высшее образование и далее
5	Возраст	*age*	указывается респондентом
6	Пол	*h5*	1 — мужской; 2 — женский
7	Удовлетворенность условиями труда	*j1.1.2*	1 — полностью удовлетворен; 2 — скорее удовлетворен; 3 — частично удовлетворен; 4 — скорее не удовлетворен; 5 — совсем не удовлетворен
8	Удовлетворенность оплатой труда	*j1.1.3*	1 — полностью удовлетворен; 2 — скорее удовлетворен; 3 — частично удовлетворен; 4 — скорее не удовлетворен; 5 — совсем не удовлетворен
9	Удовлетворенность возможностями профессионального роста	*j1.1.4*	1 — полностью удовлетворен; 2 — скорее удовлетворен; 3 — частично удовлетворен; 4 — скорее не удовлетворен; 5 — совсем не удовлетворен
10	Отраслевая принадлежность работодателя	*j4.1*	1 — легкая, пищевая промышленность; 2 — гражданское машиностроение; 3 — военно-промышленный комплекс; 4 — нефтегазовая отрасль; 5 — другая отрасль тяжелой промышленности; 6 — строительство; 7 — транспорт, связь;

[18] Вариант ответа «Все неофициально» не использован в расчетах, поскольку выходит за рамки исследуемой проблематики.

№	Показатель	Сокр.	Вариант ответа
10	Отраслевая принадлежность работодателя	*j4.1*	8 — сельское хозяйство; 9 — органы управления; 10 — образование; 11 — наука, культура; 12 — здравоохранение; 13 — армия, Министерство внутренних дел (МВД), органы безопасности; 14 — торговля, бытовое обслуживание; 15 — финансы; 16 — энергетическая промышленность; 17 — жилищно-коммунальное хозяйство (ЖКХ); 18 — операции с недвижимостью; 19 — социальное обслуживание; 20 — юриспруденция; 21 — церковь; 22 — химическая промышленность; 23 — деревообрабатывающая промышленность; 24 — спорт, туризм, развлечения; 25 — услуги населению; 26 — IT, информационные технологии; 27 — экология, защита окружающей среды; 28 — организация общественного питания; 29 — средства массовой информации (СМИ), издательство, телекоммуникации; 30 — реклама, маркетинг; 31 — общественные организации
11	Продолжительность работы на последнем месте трудоустройства, лет	*j5a*	указывается респондентом
12	Наличие подчиненных	*j6*	1 — имеются; 2 — отсутствуют
13	Продолжительность рабочей недели, часов	*j6.2*	указывается респондентом
14	Число рабочих дней за последний месяц	*j7.1*	указывается респондентом
15	Размер полученной оплаты труда по основному месту работы за последний месяц после вычета налогов	*j10*	указывается респондентом
16	Численность сотрудников организации, всего	*j13*	указывается респондентом
17	Существует ли долг организации по оплате труда перед работником	*j14*	1 — существует; 2 — отсутствует
18	Число календарных дней отпуска (отпусков в сумме)	*j21b*	указывается респондентом

Продолжение табл. 1

№	Показатель	Сокр.	Вариант ответа
19	Является ли производство вредным или опасным, дающим право на досрочное назначение трудовой пенсии, дополнительные выплаты или льготы	*j21.3*	1 — да; 2 — нет
20	Является ли государство владельцем или совладельцем организации	*j24*	1 — да; 2 — нет
21	Являются ли владельцами или совладельцами организации российские частные лица, российские частные фирмы	*j25*	1 — да; 2 — нет
22	Представьте себе не очень приятную картину: организация, где вы работаете, по каким-то причинам завтра закроется, и все работники будут уволены. Насколько вы уверены в том, что сможете найти работу не хуже той, на которой работаете сейчас?	*j22*	1 — полностью уверен; 2 — скорее уверен; 3 — частично уверен; 4 — не очень уверен; 5 — совсем не уверен
23	Насколько вас беспокоит то, что вы можете потерять работу?	*j31*	1 — очень беспокоит; 2 — значительно беспокоит; 3 — частично беспокоит; 4 — не очень беспокоит; 5 — совсем не беспокоит
24	Наличие работы по совместительству; временной работы по контракту, соглашению, договору подряда, гранту; индивидуальной работы по лицензии или без нее	*j32.1*	1 — имеется; 2 — отсутствует
25	Получение пенсии в настоящее время	*j73*	1 — имеется; 2 — отсутствует
26	Наличие кредита или намерение взять кредит в ближайшие 12 месяцев	*j200*	1 — имеется; 2 — отсутствует
27	Наличие инвалидности	*m20.7*	1 — имеется; 2 — отсутствует
28	Частота посещения врача в течение года	*l5.0*	1 — несколько раз в месяц; 2 — один раз в месяц; 3 — 2–3 раза в год; 4 — один раз в год; 5 — реже одного раза в год

Окончание табл. 1

№	Показатель	Сокр.	Вариант ответа
29	В течение последних 12 месяцев вам уменьшали зарплату или сокращали часы работы не по вашему желанию?	*j18.2*	1 — да; 2 — нет

Источник: РМЭЗ ВШЭ. URL: http://www.hse.ru/rlms (дата обращения: 11.05.2023).

В свою очередь, категориальные данные разделены на две подгруппы. К первой отнесены факторы, значения которых можно ранжировать по важности, например для переменных *j1.1.3* «Удовлетворенность оплатой труда», *j31* «Обеспокоенность возможной потерей работы» более высокое значение соответствует более низкому уровню удовлетворенности. Во вторую группу включены переменные, характеризующие принадлежность индивида к какой-либо категории, например *region* «Регион», *h5* «Пол», *j4.1* «Отрасль», *marst* «Семейное положение». Значения данных факторов невозможно ранжировать по уровню значимости, но их важность для анализа требует дополнительной систематизации. Преобразование осуществлено методом One-Hot Encoding[19], основанном на создании дополнительных бинарных признаков, отражающих принадлежность к уникальному значению.

Отметим, что после стандартизации непрерывных переменных и One-Hot кодирования номинативных выборка изменилась следующим образом: непрерывные переменные были масштабированы путем преобразования статистического распределения в формат со средним значением, равным нулю, и стандартным отклонением, равным единице. После One-Hot кодирования отдельных категориальных переменных количество факторов в модели увеличилось до 103 ед.

Фрагмент итоговой выборки представлен в табл. 2.

Важно отметить, что распределение наблюдений целевой переменной (*type_of_salary*) сильно не сбалансировано: 12 383 респондентов получают официальную заработную плату в полном объеме, и только 1184 — серую. С целью нивелирования эффекта несбалансированности для обучающей выборки был применен метод увеличения числа объектов миноритарного класса (Synthetic Minority Oversampling Technique, SMOTE) — алгоритм предварительной обработки данных, используемый для устранения дисбаланса классов, в основе которого лежит метод k-ближайших соседей. Тренировочная выборка расширена вдвое, до 18 506 наблюдений, тестовая выборка осталась без изменений.

Все модели реализованы на языке программирования Python с использованием библиотек CatBoost и Sklearn. Для выявления оптимального порога принятия решения построена прогностическая функция, пошагово меняющая порог принятия решения на 0,001 и выбирающая значение, при котором оба параметра эндогенной переменной имеют минимальный уровень ошибки на тестовой выборке.

Оценку качества модели в целом, без привязки к конкретному порогу, провели путем анализа ROC-AUC площади (Area Under Curve) под кривой ошибок (Receiver

[19] Pedregosa et al. (2011) *Scikit-learn: Machine Learning in Python*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html (дата обращения: 11.06.2023).

Таблица 2. **Фрагмент итогового набора данных, используемого в моделировании**

№	diplom	Age**	…	region_73*	region_77*	…	j4.1_32*	marst_2*
0	6	0,31	…	0	0	…	0	1
1	6	1,43	…	0	0	…	0	0
2	6	3,25	…	0	0	…	0	1
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
13565	4	0,04	…	0	0	…	0	1
13566	3	–0,46	…	0	0	…	0	1
13567	3	–0,46	…	0	0	…	0	1

* Переменные, полученные методом One-Hot Encoder.
** Стандартизированные данные.

Operating Characteristic curve). Показатели ROC-AUC для каждой из шести моделей представлены на рис. 2, матрицы ошибок — на рис. 3.

На тестовой выборке все модели, кроме K-Nearest Neighbours, предсказали более 70 % случаев. Наилучшие результаты показали нелинейные модели CatBoost и Random Forest (более 83 % правильных ответов). Менее успешными стали модель логистической регрессии и наивного байесовского классификатора (около 70 % правильных ответов). Модель K-ближайших соседей показала результат менее 70 %, в связи с этим в дальнейших расчетах она не использована.

Для соблюдения принципа единообразия в процессе выявления факторов, влияющих на прогноз в каждой из используемых моделей, применен алгоритм Шепли (рис. 4). Важно отметить, что особенностью данного алгоритма является описание влияния всех переменных на вероятность принадлежности к определенному классу с ранжированием их по вкладу в вероятность выбора класса искомой переменной. По умолчанию визуализируются 20 наиболее значимых факторов по значению Шепли.

Shap Values весомо уменьшается в переменных, приближающихся к 20-й позиции, то есть переменные после 20-го существенного значения не влияют на вероятность выбора эндогенной переменной. Кроме того, модели по-разному отражают влияние переменных на тип заработной платы. Поэтому в итоговый перечень факторов включены только те, которые встречаются в первой двадцатке в каждой модели. Выявленные таким образом факторы классифицированы по тематическим группам:

— организационно-правовые: российское резидентство собственника, размер организации;

— трудовые: размер заработной платы, количество дней отпуска;

— социальные: семейное положение, пол респондента;

— психологические: уверенность в возможности найти альтернативное место работы, удовлетворенность профессиональным ростом и условиями труда.

Так, вероятность получения индивидом серой заработной платы увеличивается, если собственником организации, в которой он трудоустроен, является резидент Российской Федерации (*j25*). Это происходит в силу склонности частного российского бизнеса к снижению размера налогов и обязательных платежей в бюд-

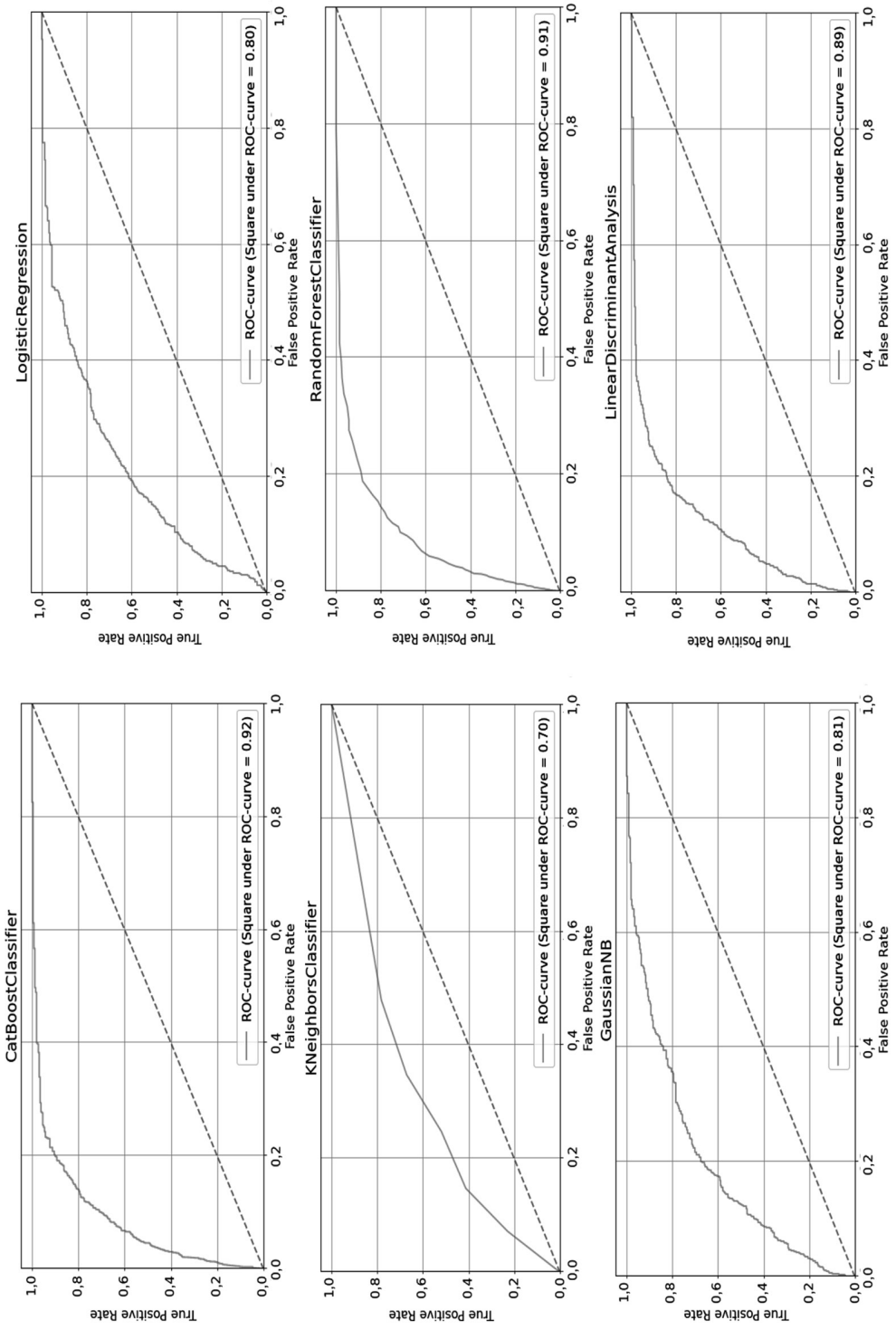

Рис. 2. Кривые ошибок (ROC) построенных моделей

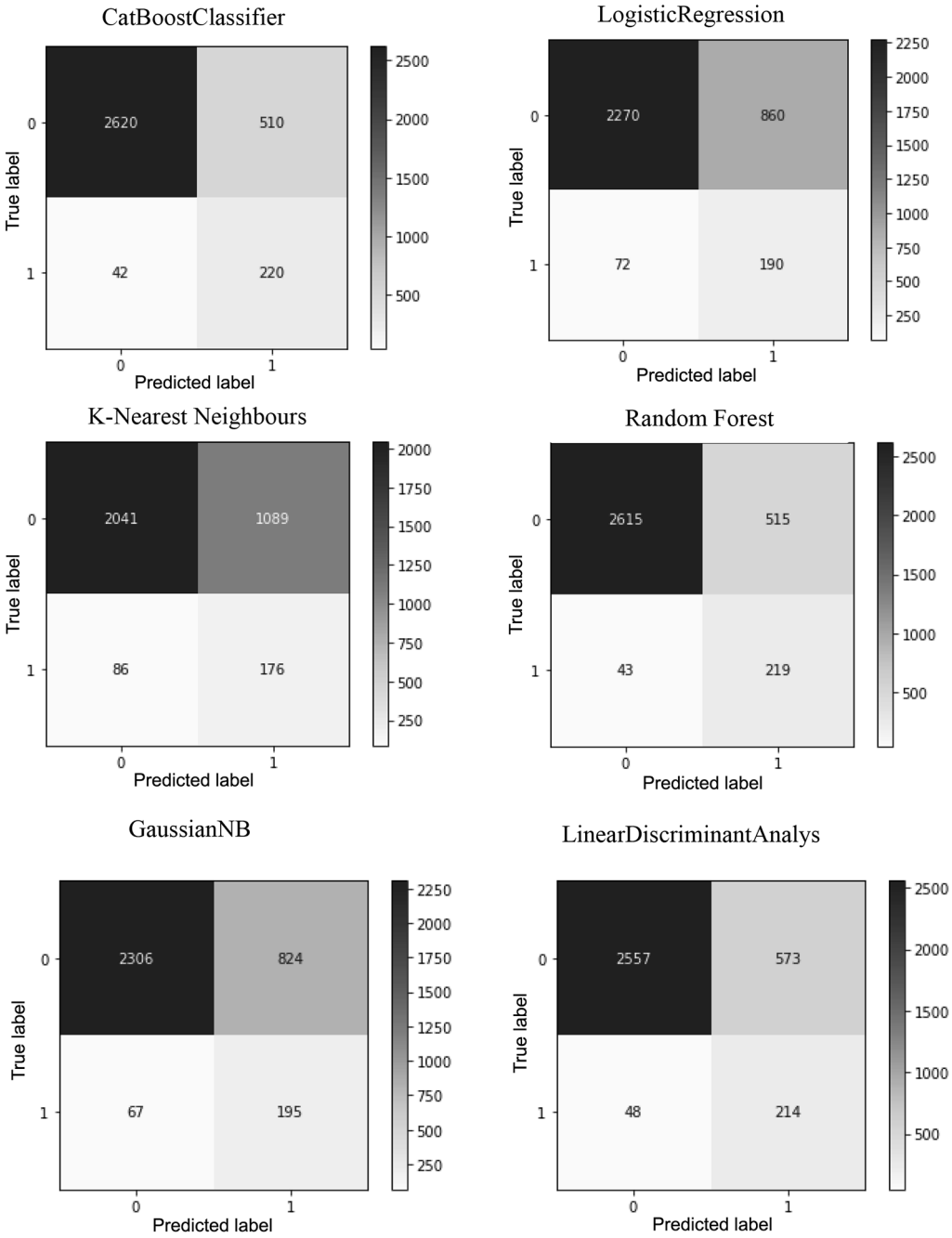

Рис. 3. Матрицы ошибок по результатам тестирования построенных моделей

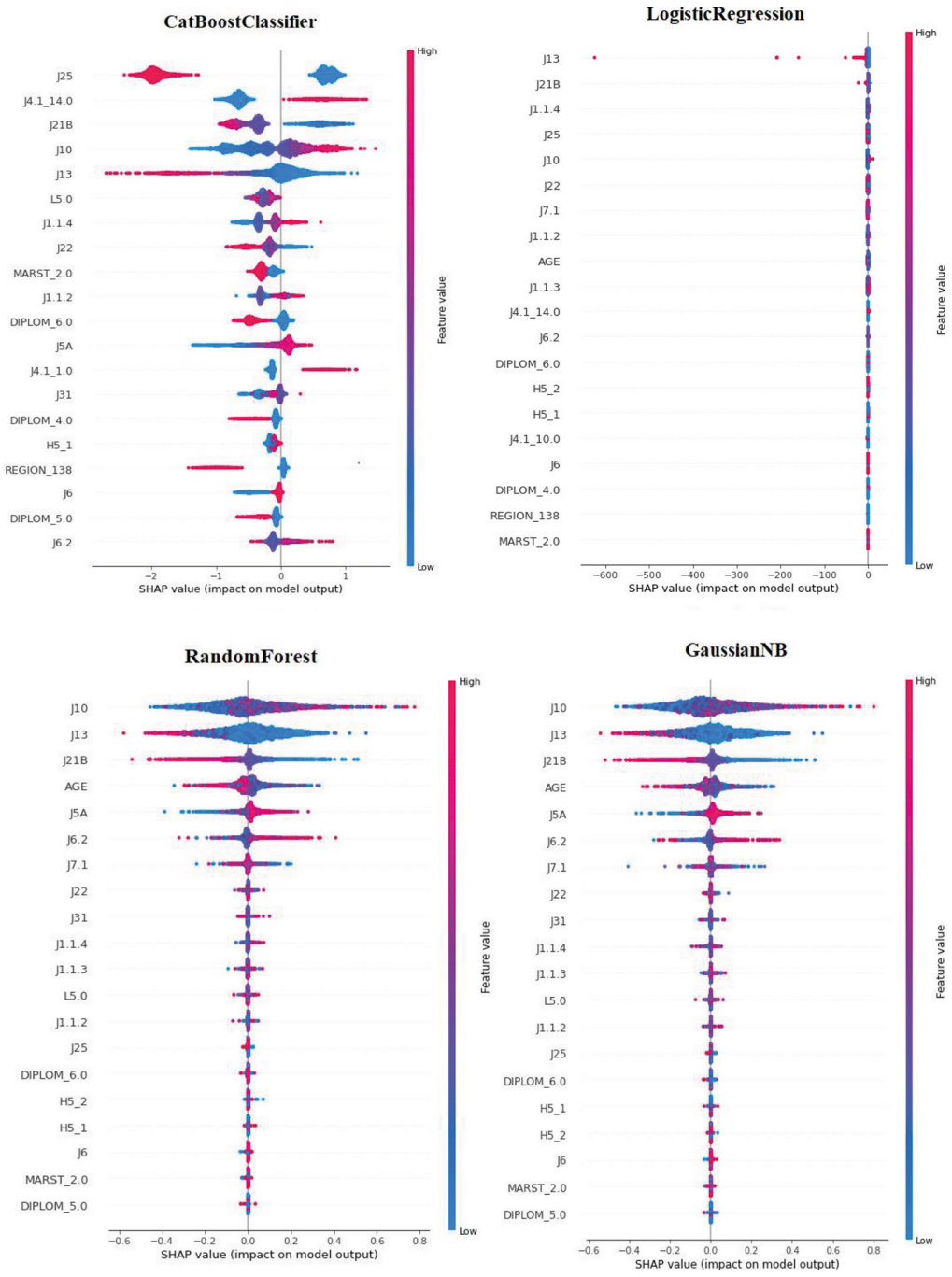

Рис. 4. Визуализация значимых факторов моделей по значению Шепли (начало рисунка)

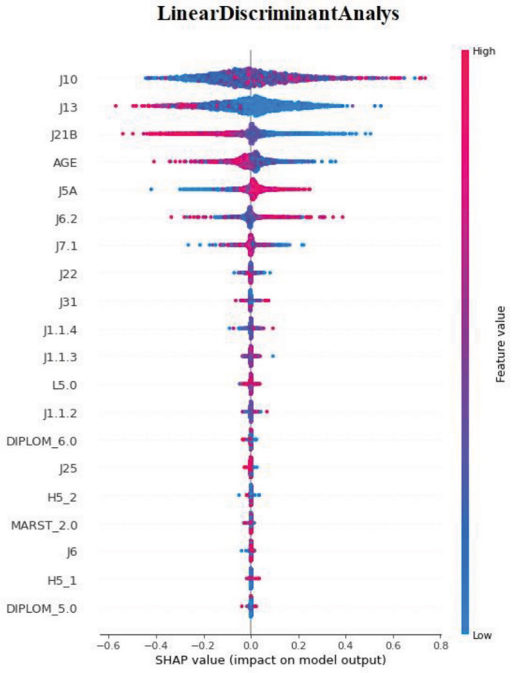

Рис. 4. Визуализация значимых факторов моделей по значению Шепли (окончание рисунка)

жет и социальные фонды. Вместе с тем предприятия с государственной формой собственности всегда отражают заработные платы сотрудников официально. Иностранные же собственники проявляют культуру социальной ответственности[20], кроме того, к ним применяется усиленный контроль со стороны различных ведомств[21].

На популярность серой зарплаты оказывает влияние и размер организации (*j13*). Модели показали, что выплачивать ее склонны небольшие организации из числа микро- и малого бизнеса, что связано с меньшим уровнем контроля со стороны надзорных органов и стремлением сэкономить на социальных взносах.

Более высокий размер заработной платы (*j10*) на фоне меньшего количества отпускных дней (*j21b*) сигнализирует о получении неофициального трудового дохода. На практике налоговая служба сравнивает уровень зарплаты на предприятии со среднеотраслевым по оценкам Росстата. Поводом для проверки может послужить ситуация, когда заявленная зарплата меньше, чем в целом по отрасли[22]. При этом привлекать внимание надзорных органов должна и ситуация, когда фактиче-

[20] Осенева, И. (2022) 'Пять причин, почему иностранные компании при уходе выплачивают 5–10 окладов'. *VC.RU.* URL: https://vc.ru/u/373586-inna-oseneva/483920-pyat-prichin-pochemu-inostrannye-kompanii-pri-uhode-vyplachivayut-5-10-okladov (дата обращения: 01.10.2023).

[21] РИА Новости. (2022) *Уходящие из России компании не нарушают права работников, заявил Роструд*. URL: https://ria.ru/20220623/rostrud-1797579762.html (дата обращения: 01.10.2023).

[22] Зайцева, М. (2022) 'Как налоговики выявляют серую зарплату: разбор эксперта'. *Klerk.ru.* URL: https://www.klerk.ru/buh/articles/538055/ (дата обращения: 01.10.2023).

ская зарплата значительно выше среднерыночной, а продолжительность отпуска не соответствует законодательно установленной.

Отдельно стоит отметить социальные факторы. Результаты моделирования показали, что вероятность получения серой заработной платы увеличивается, если речь идет об индивиде мужского пола (*h5_1*), состоящем в браке (*marst_2*). Данная ситуация обусловлена стремлением респондентов обеспечить более высокий уровень жизни членам семьи, в том числе за счет получения части зарплаты «в конверте» либо без официального трудоустройства.

Примечательно, что лица, получающие серую заработную плату, уверены, что в случае потери работы смогут найти место трудоустройства не хуже. При этом сотрудники, получающие зарплату «в конверте», не удовлетворены возможностями профессионального роста и условиями труда.

Подводя итог, отметим, практическая значимость исследования состоит в том, что полученные результаты могут лечь в основу совершенствования методов выявления серой заработной платы и оценки ее масштабов на территории России. В частности, предложенный инструментарий может быть использован Федеральной налоговой службой при проведении выездных проверок посредством выборочного анонимного анкетирования сотрудников организаций. Также данный материал актуален для подразделений Банка России и может быть применен в оценке влияния серой заработной платы на потребительскую активность населения в рамках анализа проводимой денежно-кредитной политики.

Заключение

Вопрос снижения доли серой заработной платы имеет существенное значение как на микроуровне, так и на уровне государства в целом. Наличие проблем, связанных с неформальной занятостью, выплатами «в конвертах», приводит не только к низкой собираемости налогов в бюджеты и страховых взносов в государственные внебюджетные фонды, но и к увеличению количества случаев нарушения трудовых прав работников, особенно в сферах оплаты и охраны труда, а отсутствие документально подтвержденного стажа негативно влияет на пенсионные права работников.

Для обнаружения признаков зарплаты «в конверте» в исследовании были применены методы логистической регрессии, дискриминантного анализа, градиентного бустинга, случайного леса и вероятностного байесовского классификатора. Факторы, оказавшиеся значимыми во всех моделях, объединены в четыре тематические группы. Наиболее важными оказались статус собственника компании (резидент или нерезидент), ее размер и отраслевая принадлежность, семейное положение индивида, продолжительность трудового отпуска, а также уровень тревоги по поводу потери рабочего места в случае ликвидации организации. Итоговое тестирование моделей позволило сделать выбор в пользу градиентного бустинга, так как данный алгоритм показал более качественные результаты и позволил верно определить владельцев официальной и серой заработной платы более чем в 82 % случаев.

Эффективным методом выявления зарплат «в конверте» может стать взаимодействие с линейными работниками компаний, в частности их анкетирование по вопросам, показавшим высокую степень влияния в ходе построения модели.

Отдельным направлением снижения уровня серых заработных плат Правительство РФ видит разработку пакета стимулирующих мер для работников и работодателей. Так, на стратегической сессии в правительстве обсуждалось введение мер поддержки компаний, которые платят достойную белую зарплату[23]. Кроме того, действенной мерой может стать использование больших данных для мониторинга компаний, а также проведение проверок организаций в случае обнаружения признаков получения серой заработной платы с помощью ресурсов Федеральной налоговой службы. На региональном уровне актуализируется определение приоритетных отраслей с наибольшим риском теневой занятости и установление ежегодных целевых показателей ее снижения, введение мер поддержки компаний, которые платят достойную белую зарплату.

Еще одной немаловажной мерой видится планомерное повышение минимального размера отплаты труда (МРОТ). Поскольку значительная доля работников, как было отмечено, получает МРОТ только на бумаге, а реальная плата за их труд выше с учетом выплат «в конверте», увеличение МРОТ в такой ситуации даст положительный эффект: общая сумма выплат сохранится на прежнем уровне, а налоговые поступления и платежи во внебюджетные фонды возрастут. Несмотря на то что вслед за ростом минимального размера оплаты труда увеличится общая сумма заработных плат работников бюджетной сферы, общий экономический эффект будет вполне сопоставим с дополнительными расходами бюджетов.

Полагаем, что рассмотренные меры в совокупности позволят значительно снизить востребованность серой заработной платы, привлечь нарушителей к ответственности, а главное — обеспечить права работника на достойное официальное денежное вознаграждение за труд.

Литература

Абзалилова, Л. Р. (2021) 'Моделирование оттока кадров в крупной компании с применением технологий интеллектуального анализа данных', *Экономика и управление: научно-практический журнал,* 3 (159), с. 152–157.

Вередюк, О. В. (2020) 'Динамика субъективного благополучия при внутрифирменной трудовой мобильности в России', *Мониторинг общественного мнения: экономические и социальные перемены,* 1, с. 391–407.

Гавриленко, Ю. Е. (2022) 'Методы устойчивой кластеризации регионов России по занятости населения', *Федерализм*, Т. 27, 3 (107), с. 160–177.

Гимпельсон, В. Е. (2021) *Зарплата и потоки на российском рынке труда в условиях ковида*. М.: Изд. дом Высшей школы экономики.

Гимпельсон, В. Е. и Капелюшников, Р. И. (2020) *Российский рынок труда через призму демографии.* М.: Изд. дом Высшей школы экономики.

Гимпельсон, В. Е., Капелюшников, Р. И. и Шарунина, А. В. (2018) 'Низкооплачиваемые рабочие места на российском рынке труда: есть ли выход и куда он ведет?', *Экономический журнал ВШЭ,* 22 (4), с. 489–530.

Гимпельсон, В. Е., Капелюшников, Р. И., Лукьянова А., Рыжикова, З. и Куляева, Г. (2010) 'Формы собственности в России: различия в заработной плате', *Журнал Новой экономической ассоциации,* 5, с. 48–72.

Дембовский, И. А. и Машков, А. А. (2019) 'Оценка доли серых зарплат в регионах России для агентного моделирования процессов трудоустройства жителей', в Иванцова Е. Н., Уваров В. М.

[23] Виноградова, Е. (2023) 'Власти обсудили, как сократить число получающих зарплату «в конверте»'. *РБК*. URL: https://www.rbc.ru/economics/28/04/2023/644a69299a79470b33e7efe1 (дата обращения: 01.07.2023).

(сост.) *Сборник докладов XII Междунар. науч.-практич. конф. студентов, аспирантов и молодых ученых. В 3 т. Т. 3*. Старый Оскол, 2019, с. 177–181.
Журавлева, Т. Л. (2015) 'Платит ли российское государство «справедливую» зарплату: обзор исследований', *Вопросы экономики,* 11, с. 62–85.
Зарова, Е. В. и Дубравская, Э. И. (2020) 'Метод «случайный лес» в исследовании влияния макроэкономических показателей регионального развития на уровень неформальной занятости', *Вопросы статистики*, 27 (6), с. 37–55.
Ляхнова, М. В. и Рудаев, Г. С. (2021) 'Оценка социально-экономических факторов, влияющих на желание сменить работу', *Modern Science*, 3-2, с. 102–109.
Мальцева, А. В., Махныткина, О. В. и Шилкина, Н. Е. (2015) 'Многомерные классификационные модели в сравнительном анализе социально-структурной специфики регионов России', *Науковедение,* 7 (6). URL: http://naukovedenie.ru/PDF/13EVN615.pdf (дата обращения: 30.10.2023).
Хохлова, О. А., Хохлова, А. Н. и Чойжалсанова, А. Ц. (2022) 'Разработка алгоритма анализа вакансий на рынке труда по данным из открытых источников', *Вопросы статистики*, 29 (4), с. 33–41.
Шарунина, А. В. (2013) 'Является ли российский «бюджетник» неудачником? Анализ межсекторных различий в оплате труда', *Экономический журнал ВШЭ*, Т. 17, 1, с. 75–107.
Dorogush, A. V., Ershov, V. and Gulin, A. (2017) 'CatBoost: Gradient boosting with categorical features support', *Workshop on ML Systems at NIPS 2017*. URL: http://learningsys.org/nips17/assets/papers/paper_11.pdf (дата обращения: 26.06.2023).
Geanderson, E., Figueiredo, E., Veloso, A., Viggiato, M. and Ziviani, N. (2020) 'Understanding learning software defect predictions', *Automated Software Development*, 27, pp. 369–392.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python, *JMLR*, 12, pp. 2825–2830.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin, A. (2019) 'CatBoost: Unbiased boosting with categorical features', *arXiv: 1706.09516.* URL: https://arxiv.org/pdf/1706.09516.pdf (дата обращения: 20.06.2023).
Valouet, T., Al-Memar, M., Fourie, H., Bobdiwala, S., Saso, S., Pipi, M., Stalder, C., Bennett, P., Timmerman, D., Bourte, T. and De Moor, B. (2022) 'Gradient boosted trees with individual explanations: An alternative to logistic regression for predicting viability in the first trimester of pregnancy', *Computer Methods and Programs in Biomedicine*, 213, 106520.

Статья поступила в редакцию: 28.07.2023
Статья рекомендована к печати: 16.05.2024

Контактная информация:

Куницына Наталья Николаевна — д-р экон. наук, проф.; https://orcid.org/0000-0001-9336-8100, nkunitcyna@ncfu.ru
Метель Юрий Андреевич — канд. экон. наук; https://orcid.org/0009-0001-5334-0020, akademik.st.2018@mail.ru

Modeling the influence of social factors on the dynamics of gray wages

N. N. Kunitsyna [1], Yu. A. Metel [2]

[1] North Caucasus Federal University,
1, ul. Pushkina, Stavropol, 355017, Russian Federation
[2] Stavropol Territorial Division of the Southern Main Branch of the Central Bank of the Russian Federation,
286, ul. Lenina, Stavropol, 355035, Russian Federation

For citation: Kunitsyna, N. N. and Metel, Yu. A. (2024) 'Modeling the influence of social factors on the dynamics of gray wages', *St. Petersburg University Journal of Economic Studies,* 40 (3), pp. 460–482. https://doi.org/10.21638/spbu05.2024.306 (In Russian)

The problem of whitewashing wages of a significant share of Russians escalated during the pandemic, post-covid recovery and, especially, against the background of geopolitical shocks.

The unprecedented expansion of state support measures provided targeted and officially has prompted entrepreneurs and citizens to get out of the shadows. Despite the obvious positive changes, the problem of salaries "in envelopes" has not eradicated itself: in modern publications, scientists return to it repeatedly. When assessing this socio-economic phenomenon, science relies mainly on methods of regression, panel analysis, instrumental variables, etc. In this work to assess the influence of social factors on the dynamics of gray wages we used the gradient boosting method, linear modeling, random forest and a naive Bayesian classifier. As the initial information, we used the results of the surveys of the Russian Monitoring of the Economic Situation and Health of the Population by Higher School of Economics, which are freely available. The result of the simulation is the dependence of the receiving wages method (official, "in an envelope") on a number of factors. The organization's industry affiliation, size, ownership, employee's education, duration of vacation, satisfaction with professional growth and working conditions have the strongest influence. The applied value of the results obtained is the possibility of generating control effects, both at the state level and at the level of business entities in the direction of leveling the identified reasons for receiving gray salaries and reducing the scale of the hidden wage fund.

Keywords: salary, gray salary, gradient booster, linear models, random forest.

References

Abzalilova, L. R. (2021) 'Modeling the Outflow of Personnel in a Large Company Using Data Mining Technologies', *Economics and Management: A Scientific and Practical Journal,* 3 (159), pp. 152–157. (In Russian)

Dembovsky, I. A. and Mashkov, A. A. (2019) 'Estimation of the share of gray wages in the regions of Russia for agent modeling of the processes of employment of residents', in Ivantsova E. N., Uvarov V. M. (comp.) *Sbornik dokladov XII Mezhdunarodnoi nauchno-prakticheskoi konferentsii studentov, aspirantov i molodykh uchenykh. In 3 vols, Vol. 3.* Stary Oskol, 2019, pp. 177–181. (In Russian)

Dorogush, A. V., Ershov, V. and Gulin, A. (2017) 'CatBoost: Gradient enhancement with support for categorical functions', *Seminar on ML systems at NIPS*. Available at: http://learningsys.org/nips17/assets/papers/paper_11.pdf (accessed: 26.06.2023).

Gavrilenko, I. E. (2022) 'Methods of sustainable clustering of Russian Regions by employment', *Federalizm,* 27, 3 (107), pp. 160–177. (In Russian)

Geanderson, E., Figueiredo, E., Veloso, A., Viggiato, M. and Ziviani, N. (2020) 'Understanding learning software defect predictions', *Automated Software Development*, 27, pp. 369–392.

Gimpelson, V. E. (2021) *Wages and flows in the Russian labor market in the conditions of COVID.* Moscow: Publishing House HSE University. (In Russian)

Gimpelson, V. E. and Kapelyushnikov, R. I. (2020) *The Russian labor market through the prism of demography.* Moscow: Publishing House HSE University. (In Russian)

Gimpelson, V. E., Kapelyushnikov, R. I. and Sharunina, A. V. (2018) 'Low-paid jobs in the Russian labor market: Is there a way out and where does it lead?', *HSE Economic Journal*, 22 (4), pp. 489–530. (In Russian)

Gimpelson, V. E., Kapelyushnikov, R. I., Lukyanova, A., Ryzhikova, Z. and Kulyaeva, G. (2010) 'Forms of Ownership in Russia: Wage differences', *Journal of the New Economic Association*, 5. pp. 48–72. (In Russian)

Khokhlova, O. A., Khokhlova, A. N. and Choyzhalsanova, A. T. (2022) 'Development of an Algorithm to Analyze Vacancies in the Labor Market Based on Open-Source Data', *Voprosy Statistiki,* 29 (4), pp. 33–41. (In Russian)

Lyakhnova, M. V. and Rudaev, G. S. (2021) 'Assessment of socio-economic factors affecting the desire to change jobs', *Modern Science*, 3-2, pp. 102–109. (In Russian)

Mal'tseva, A. V., Makhnytkina, O. V., and Shilkina, N. E. (2015) 'Multidimensional classification models for comparative analysis of Russian Federation regions social and economical specifics', *Naukovedenie,* 7 (6). Available at: http://naukovedenie.ru/PDF/13EVN615.pdf (accessed: 30.10.2023). (In Russian)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) 'Scikit-learn: Machine Learning in Python', *JMLR*, 12, pp. 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. and Gulin, A. (2019) 'CatBoost: Unbiased boosting with categorical features', *arXiv: 1706.09516*. Available at: https://arxiv.org/pdf/1706.09516.pdf (accessed: 20.06.2023).
Sharunina, A. V. (2013) 'Is the Russian "state employee" a loser? Analysis of Intersectoral Wage Differences', *HSE Economic Journal*, 17 (1), pp. 75–107. (In Russian)
Valouet, T., Al-Memar, M., Fourie, H., Bobdiwala, S., Saso, S., Pipi, M., Stalder, C., Bennett, P., Timmerman, D., Bourte, T. and De Moor, B. (2022) 'Gradient boosted trees with individual explanations: An alternative to logistic regression for predicting viability in the first trimester of pregnancy', *Computer Methods and Programs in Biomedicine*, 213, 106520.
Veredyuk, O. V. (2020) 'Internal Labor Mobility and Subjective WellBeing in Russia', *Monitoring of Public Opinion: Economic and Social Changes*, 1. pp. 391–407. (In Russian)
Zarova, E. V. and Dubravskaya, E. I. (2020) 'The Random Forest Method in Research of Impact of Macroeconomic Indicators of Regional Development on Informal Employment Rate', *Voprosy Statistiki,* 27 (6), pp. 37–55. (In Russian)
Zhuravleva, T. L. (2015) 'Does the Russian state pay a "fair" salary: A review of research', *Voprosy ekonomiki*, 11, pp. 62–85. (In Russian)

Received: 28.07.2023
Accepted: 16.05.2024

Author's information:

Natalia N. Kunitsyna — Dr. Sci. in Economics, Professor; https://orcid.org/0000-0001-9336-8100, nkunitcyna@ncfu.ru

Yuri A. Metel — PhD in Economics; https://orcid.org/0009-0001-5334-0020, akademik.st.2018@mail.ru