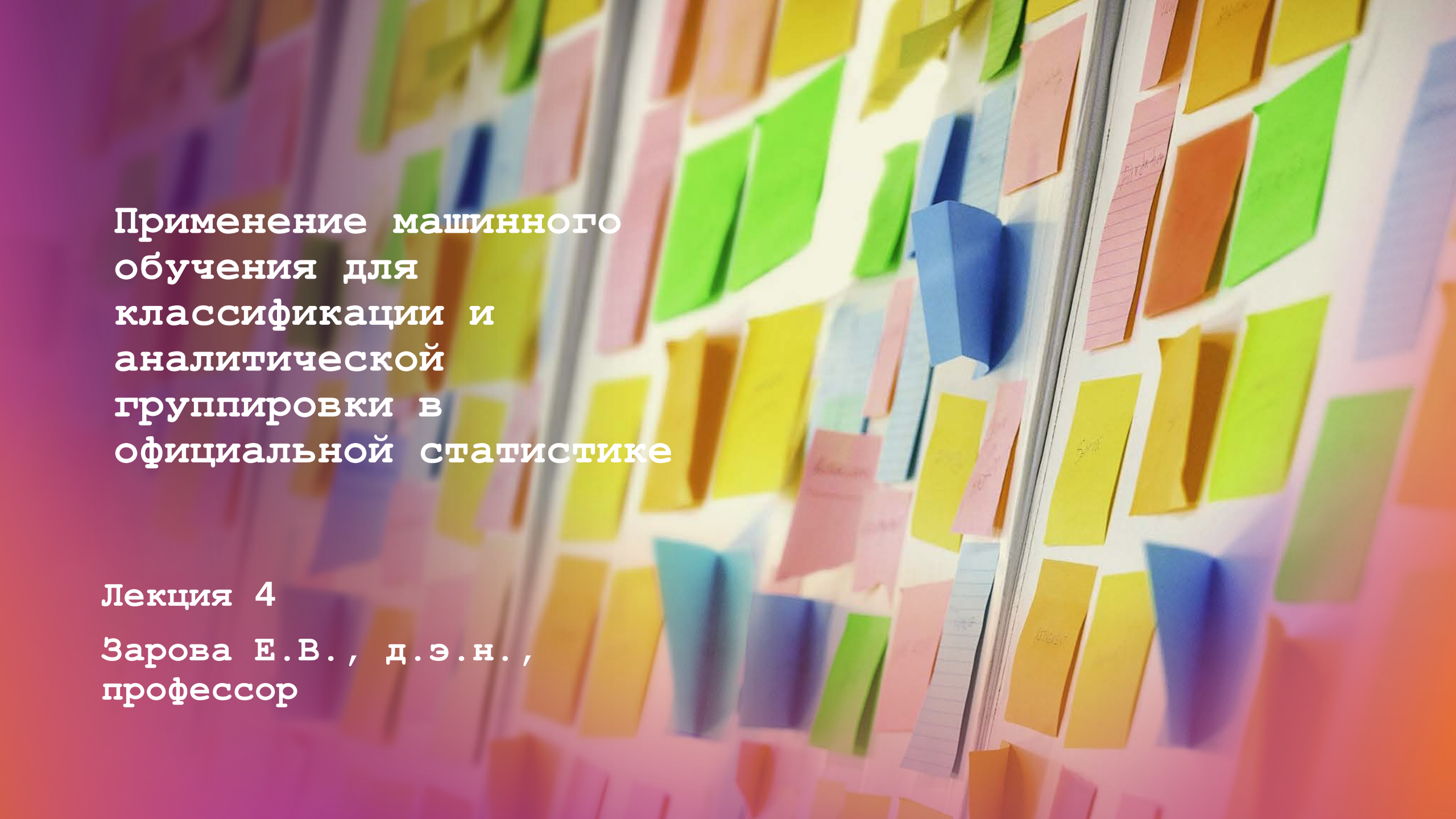


Применение машинного обучения для классификации и аналитической группировки в официальной статистике

Лекция 4

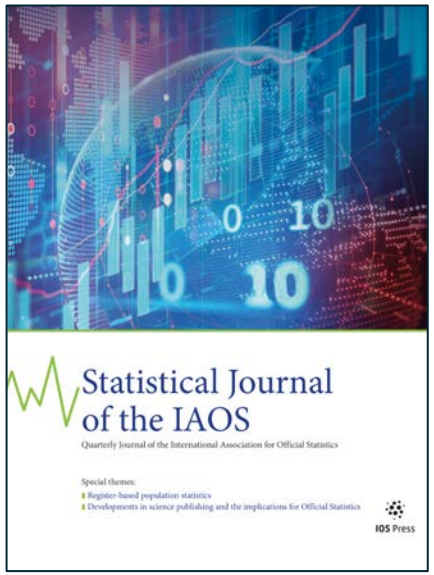
Зарова Е.В., д.э.н., профессор

План лекции

- **Классификация и многофакторная регрессия с применением методов машинного обучения и технологий нейронных сетей. Задачи, алгоритмы, примеры использования для разработки выходной информации в официальной статистике**
- **Теоретические основы: обучение с учителем, метрики качества моделей**
- **Практические кейсы из национальных статистических служб**

Актуальность темы

International Statistical Institute



Стефан Швайнфест*, директор Статистического отдела ООН: ИНТЕРВЬЮ. https://officialstatistics.com/news-blog/good-data-are-used-data-interview-stefan-schweinfest

«ХОРОШИЕ ДАННЫЕ»

«Хорошие данные» – методологически надежные»

«Хорошие данные» – используемые данные»

ВОПРОСЫ:

1. **Неиспользуемые стат. данные – «нехорошие»?**
2. **Как оценить использование публикуемых статистических данных (достаточно ли цитирования)?**

*«*Мы не можем просто предоставлять "силосы" с нашими экономическими данными, нашими экологическими данными, нашими социальными данными... Важный вопрос заключается в том, как эти явления взаимосвязаны.»*

«Силосы данных» (Data Silos) — это метафора, описывающая ситуацию, когда массивы информации (данные) хранятся в изолированных, не связанных друг с другом системах или ведомствах, подобно тому, как зерно хранится в отдельных силосных башнях.

Системообразующая роль государственной статистики

- *"Государственная статистика играет системообразующую роль в жизни общества, являясь*
  - *основой для принятия эффективных управленческих решений,*
  - *проведения анализа и прогнозирования социально-экономических процессов,*
  - *обеспечения научных исследований,*
  - *а также для информирования общества о состоянии экономики и социальной сферы."*

«Активная роль» официальной статистики

«Предвосхищающая аналитика» — это подход в официальной статистике, при котором с помощью анализа больших данных и методов искусственного интеллекта заблаговременно выявляются скрытые тенденции, риски и будущие информационные потребности, позволяя формировать статистические продукты до поступления прямых запросов от пользователей.

Машинное обучение и технологии ИИ позволяют находить неочевидные, не заложенные в традиционные методологии связи между данными, что даёт возможность не просто описывать прошлое, а предсказывать новые информационные потребности и формировать статистику будущего.

Пример международных рекомендаций для практической статистики

Машинное обучение: автоматизация обработки данных и менее трудоемкий контроль качества

International Labour Organization

21st International Conference of Labour Statisticians
11-20 October 2023

Labour market modelling

Agenda items
Econometric modelling of estimates (para. 214-218)
Nowcasting techniques and ILO Monitor in COVID-19 times (para. 21

International Labour Organization

Future work

- Policy simulation: new topics and methodological refinement
- Machine Learning: automation of data processing and less labour-intensive quality control
- Communication: reaching key stakeholders and policy makers
- Assess possibilities in skills, productivity, job quality, forced labour and 19th ICLS employment

Случайный лес

Машинное обучение в официальной статистике: ансамбли деревьев регрессии и классификации для повышения информативности данных

Случайный лес - ансамбль решающих деревьев

«Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается»

Алгоритм случайного леса (**Random Forest**) — универсальный алгоритм машинного обучения, предложенный Лео Брейманом и Адель Катлер (США, начало 2000-х годов)

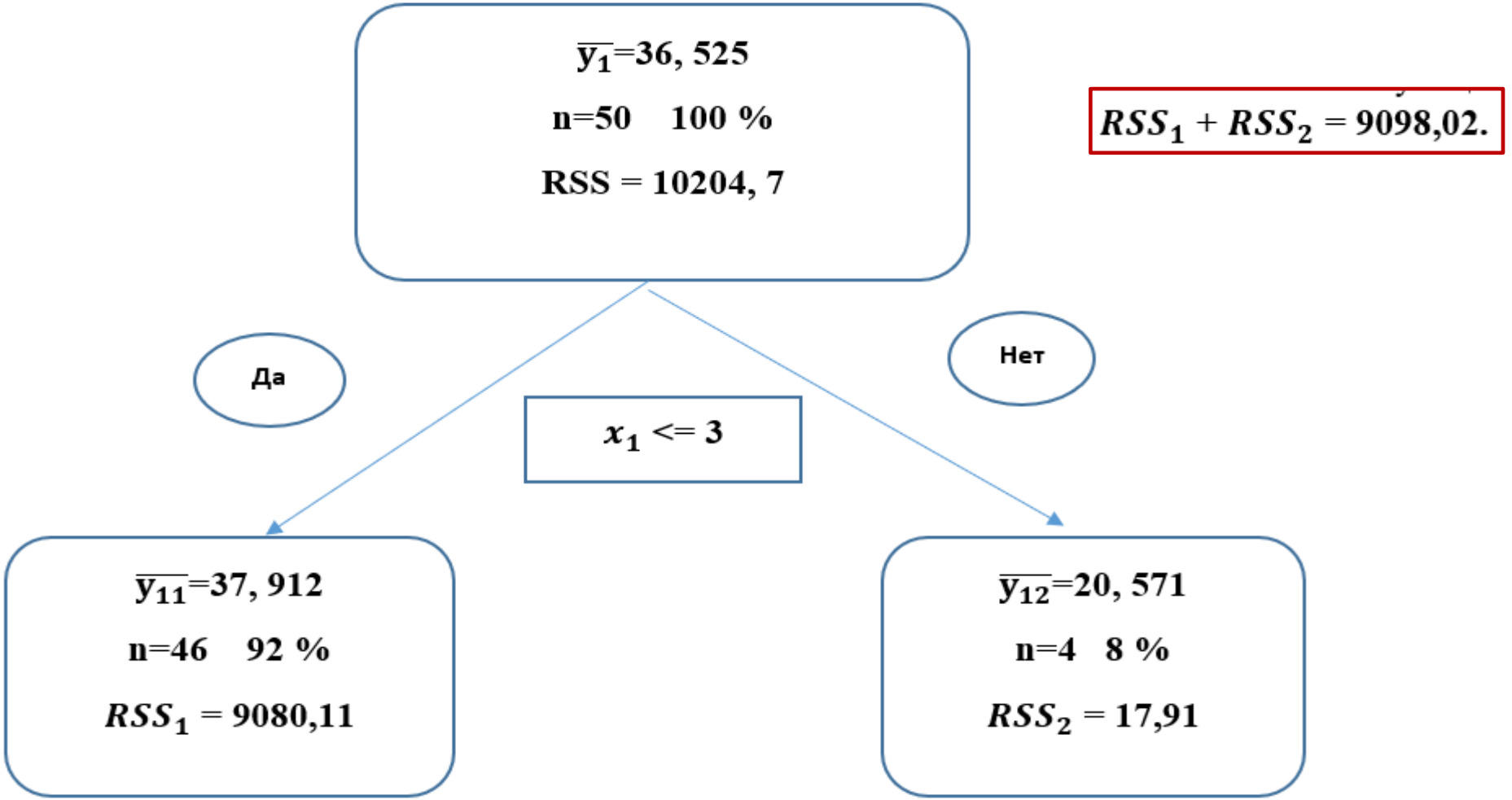
Задачи, решаемые в области машинного обучения

- Классификация
- Регрессия
- Кластеризация
- Отбор признаков
- поиск выбросов/аномалий

Основные термины, используемые в методе построения дерева решений

Объект	Наблюдение, единица совокупности, образец
Атрибут	Признак, независимая переменная
Целевая переменная	Зависимая переменная: количественная (отклик) или качественная (метка класса)
Узел	Внутренний узел дерева, узел проверки
Корневой узел (корневая вершина)	Начальный узел дерева решений
Терминальный узел (терминальная вершина), лист	Конечный узел дерева, узел решения — узел без исходящих связей
Решающее правило	Условие в узле, проверка
Критерии качества построения дерева	Для деревьев регрессий: ➢ RSS - residual sum of squares —сумма квадратов остатков. Для деревьев классификации: ➢ Частота ошибок классификации: доля обучающих наблюдений в соответствующей области, которые не принадлежат к наиболее распространенному классу: ➢ Индекс Джини - мера общей дисперсии во всех *K* классах. ➢ Энтропия

Пример 1. Дерево регрессии зависимой переменной «Среднедушевой денежный доход в расчете на одного члена домохозяйства за месяц (DOXODN), тыс. руб.» с решающим правилом разделения домохозяйств по предиктору 1 на две группы по числу членов: меньше или равно 3-м. более 3-х



$RSS_1 + RSS_2 = 9098{,}02.$

Как выбирается наилучшее ветвление?

RSS1 + RSS2 в оценке ветвления дерева — это ключевой показатель, используемый в алгоритмах построения дерева решений для задач регрессии.

RSS расшифровывается как **Residual Sum of Squares** (сумма квадратов остатков) и вычисляется как:
где:
y_i — реальное значение целевой переменной,
$\hat{y}_i$ — предсказанное значение (например, среднее значение по листу дерева).
Что происходит при ветвлении?
До ветвления у нас есть один узел (лист) с некоторым набором наблюдений. Мы вычисляем **RSS_parent** для этого узла.
При ветвлении узел разбивается на два дочерних узла по некоторому признаку и порогу.
Для каждого дочернего узла вычисляется своя сумма квадратов остатков: **RSS1** (для левой ветви) и **RSS2** (для правой ветви).
RSS1 + RSS2 — это общая ошибка модели **после** ветвления.

Как это используется для оценки ветвления?
Алгоритм выбирает то разбиение, которое **максимально уменьшает RSS**, то есть минимизирует величину:

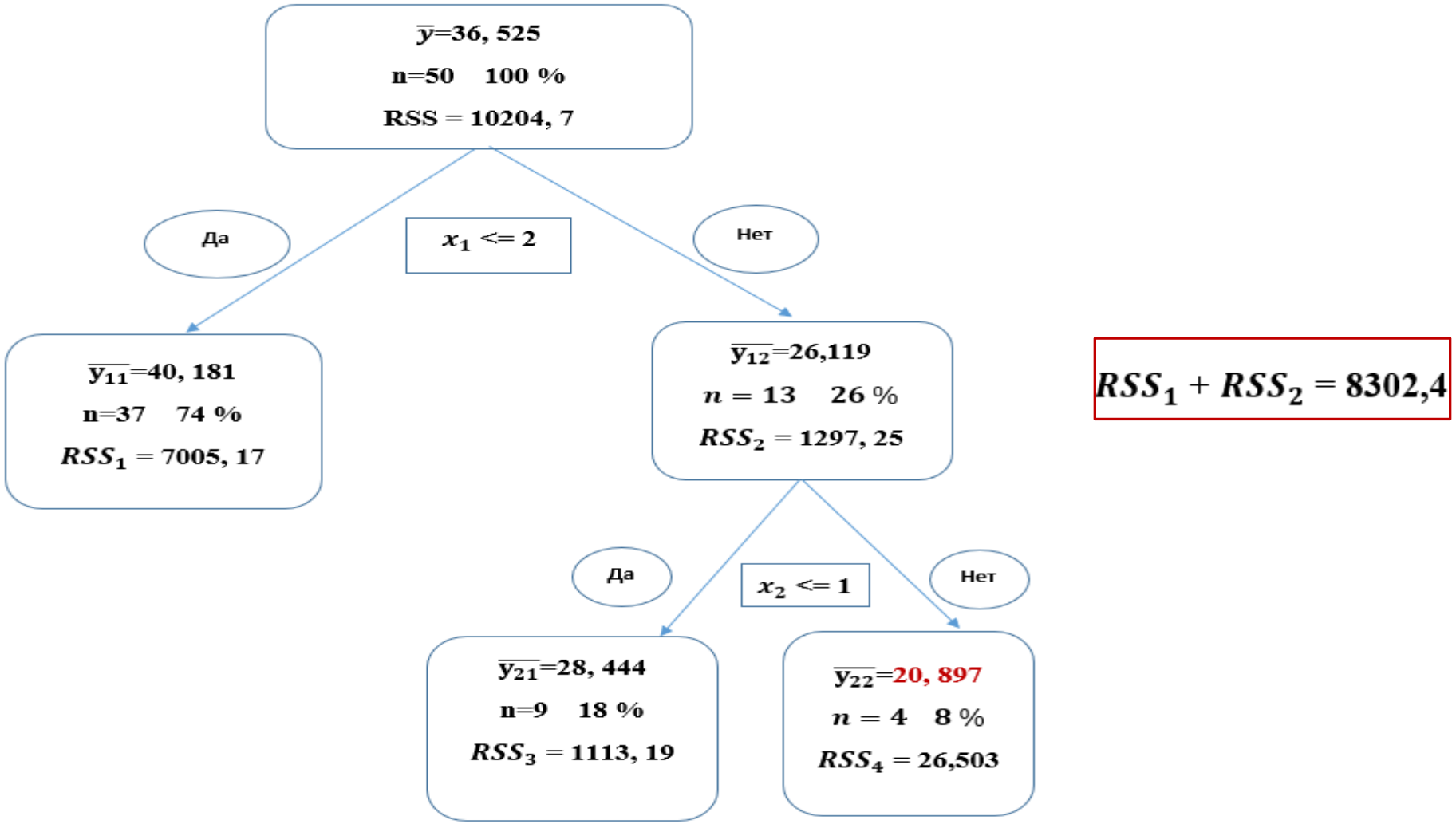
$$RSS1 + RSS2$$

Чем **меньше** сумма RSS1 + RSS2, тем лучше разделились данные, и тем более однородными стали дочерние узлы.
Процесс выбора лучшего разбиения:
Для всех возможных признаков и порогов разбиения вычисляется:

$$\Delta RSS = RSS_{parent} - (RSS1 + RSS2)$$

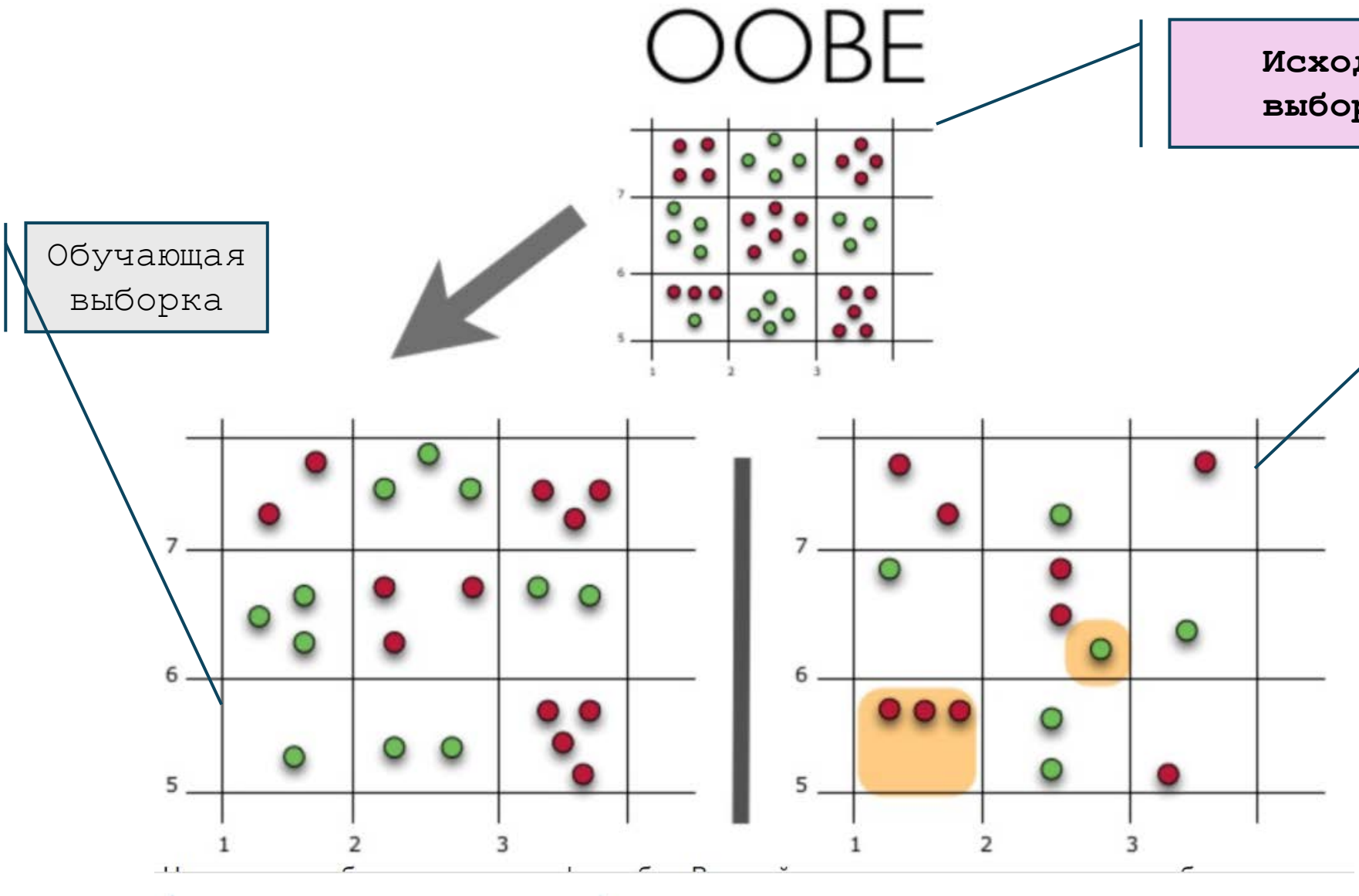
Разбиение с **максимальным** ΔRSS (наибольшим уменьшением ошибки) выбирается как лучшее.

Дерево регрессии зависимой переменной «Среднедушевой денежный доход в расчете на одного члена домохозяйства за месяц (DOXODN), тыс. руб.» с решающим правилом разделения домохозяйств по предиктору 1 на две группы по числу членов: меньше или равно 2-м, более 2-х.



$$RSS_1 + RSS_2 = 8302{,}4$$

Out-of-bag error (OOBE) – оценка модели на тестовой совокупности



Доля верных ответов классификатора =

$$\frac{11}{15} * 100\% = 73.33\%$$

Построение ансамбля деревьев решений

$$\varepsilon_i(x) = b_i(x) - y(x), i = 1, \ldots, n$$

$$a(x) = \frac{1}{n}\sum_{i=1}^{n} b_i(x)$$

$$\mathbb{E}_x\left[(b_i(x) - y(x))^2\right] = \mathbb{E}_x\left[\varepsilon_i^2(x)\right]$$

«Белый шум», MIN

$$\begin{aligned}
\mathbb{E}_n &= \mathbb{E}_x\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} b_i(x) - y(x)\right)^2 \\
&= \mathbb{E}_x\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \varepsilon_i\right)^2 \\
&= \tfrac{1}{n^2}\mathbb{E}_x\left(\textstyle\sum_{i=1}^{n} \varepsilon_i^2(x) + \sum_{i\neq j} \varepsilon_i(x)\varepsilon_j(x)\right) \\
&= \tfrac{1}{n}\mathbb{E}_1
\end{aligned}$$

«Белый шум», MIN

Алгоритм построения случайного леса, состоящего из N деревьев

Для каждого $n = 1, \ldots, N$:

- Сгенерировать выборку X_n с помощью бутстрэпа;
- Построить решающее дерево b_n по выборке X_n:
  — по заданному критерию мы выбираем лучший признак, делаем разбиение в дереве по нему и так до исчерпания выборки
  — дерево строится, пока в каждом листе не более $n_{\min}$ объектов или пока не достигнем определенной высоты дерева
  — при каждом разбиении сначала выбирается m случайных признаков из n исходных, и оптимальное разделение выборки ищется только среди них.

Итоговый классификатор $a(x) = \frac{1}{N} \sum_{i=1}^{N} b_i(x)$,

Для задачи классификации решение принимается «голосованием» по большинству

в задаче регрессии — по среднему значению.

Пример 2. Выявление факторов, влияющих на Среднедушевой расход на конечное потребление (руб.) в микроданных ОБДХ

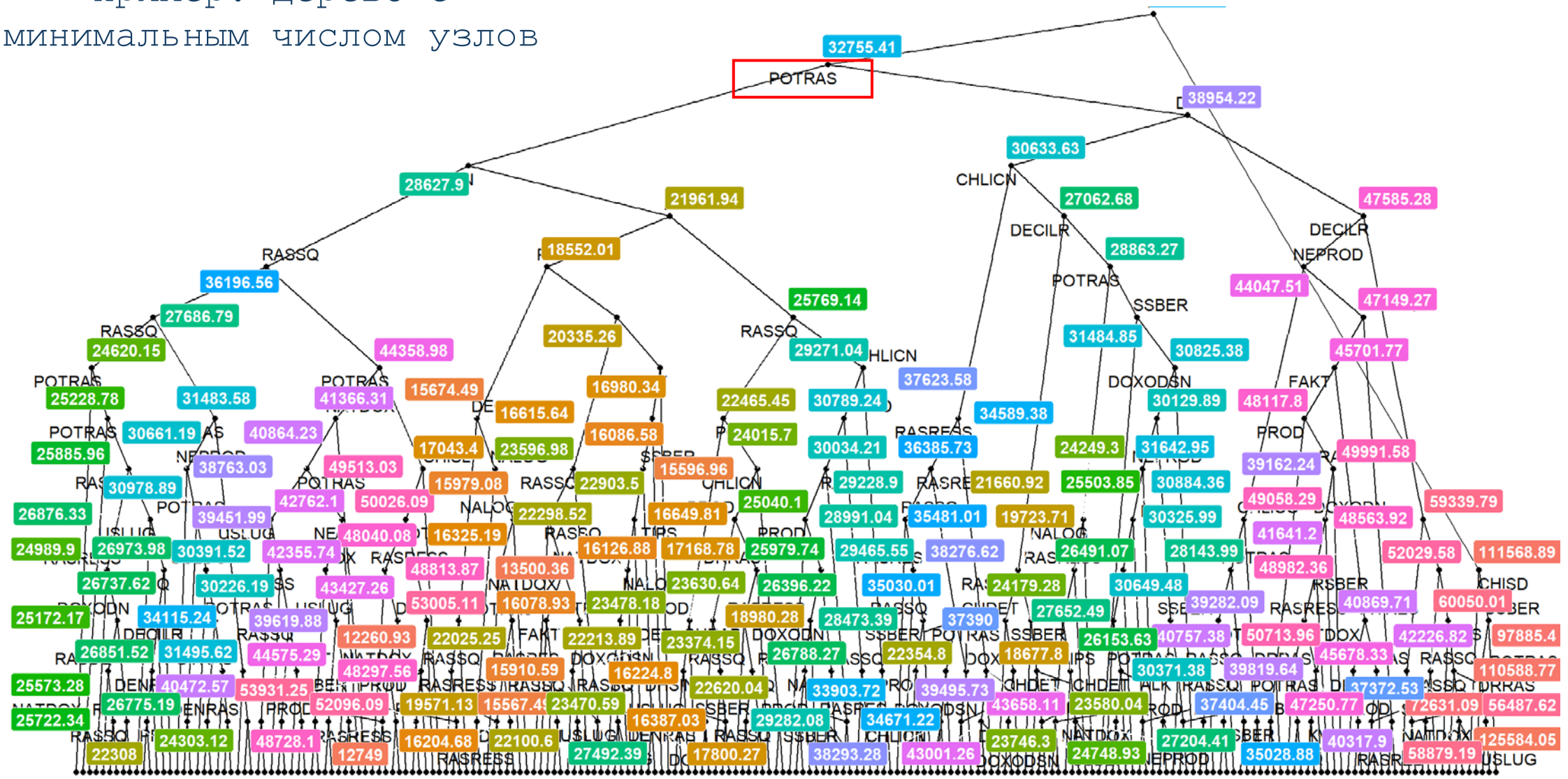
Основные расчетные показатели (файл FC)

Имя переменной	Описание переменной
PER	Период разработки
TER	Шифр территории
MEST	Тип населенного пункта
BUD	Номер бюджета
CHLICO	Число лиц в домохозяйстве
CHLICN	Число наличных лиц в домохозяйстве
CHISL	Группировка по числу наличных лиц в домохозяйстве (5 гр
CHDET	Фактическое число детей до 16 лет
CHISD	Группировка по числу детей до 16 лет в домохозяйстве (5 г
CLDP	Число человеко-дней питания
DOXODSN	Денежный доход
DOXODN	Среднедушевой денежный доход
RASRESS	Располагаемые ресурсы
RASRES	Среднедушевые располагаемые ресурсы
DECILR	Дециль по среднедушевым располагаемым ресурсам
RASSQ	Расход на конечное потребление
RASQ	Среднедушевой расход на конечное потребление

Целевая переменная

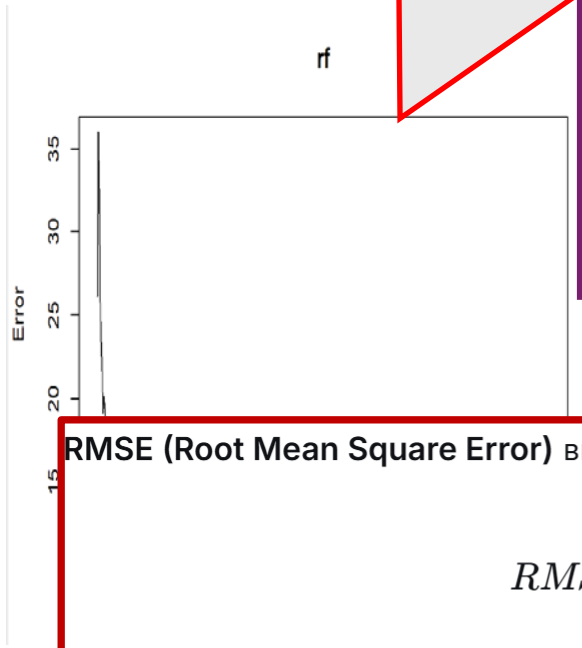
POTRAS	Потребительские расходы
PROD	Расходы на покупку продуктов питания
PITRES	Расходы на питание вне дома
ALK	Расходы на покупку алкогольных напитков
NEPROD	Расходы на покупку непродовольственных товаров
USLUG	Расходы на оплату услуг
PROMPOT	Расходы на промежуточное потребление и валовое накопление
NALOG	Налоги, сборы, платежи
DRRAS	Другие расходы
DENRAS	Денежные расходы
FAKT	Прирост финансовых активов
NATDOX	Натуральные поступления
R1GV91	Субсидии на оплату жилья и коммунальных услуг, получаемые перечислением на банковские счета
R1GV92	Субсидии на оплату жилья и коммунальных услуг, включенные в счета на оплату
R1GV93	Иные денежные компенсации на оплату жилья и коммунальных услуг
R1GV94	Льготы (скидки) при оплате счетов за электричество
R1GV95	Льготы (скидки) при оплате счетов за услуги телефонной сети
R1GV96	Льготы (скидки) на покупку топлива
R1GV97	Льготы по оплате за вывоз мусора и обеззараживание твердых бытовых отходов
R1GV98	Единовременная помощь на ремонт жилья и др.

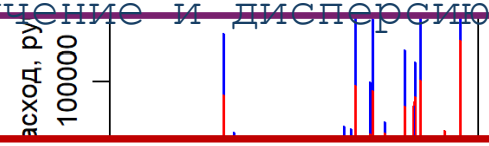
Пример: дерево с минимальным числом узлов



Финальный RMSE при построении ансамбля деревьев регрессии — это итоговый показатель ошибки модели, полученный после агрегации предсказаний всех деревьев в ансамбле

Финальный RMSE обычно значительно **ниже**, чем у одного дерева, благодаря усреднению предсказаний множества независимых деревьев, что снижает переобучение и дисперсию.





RMSE (Root Mean Square Error) вычисляется по формуле:

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{y}_j^{\text{ensemble}})^2}$$

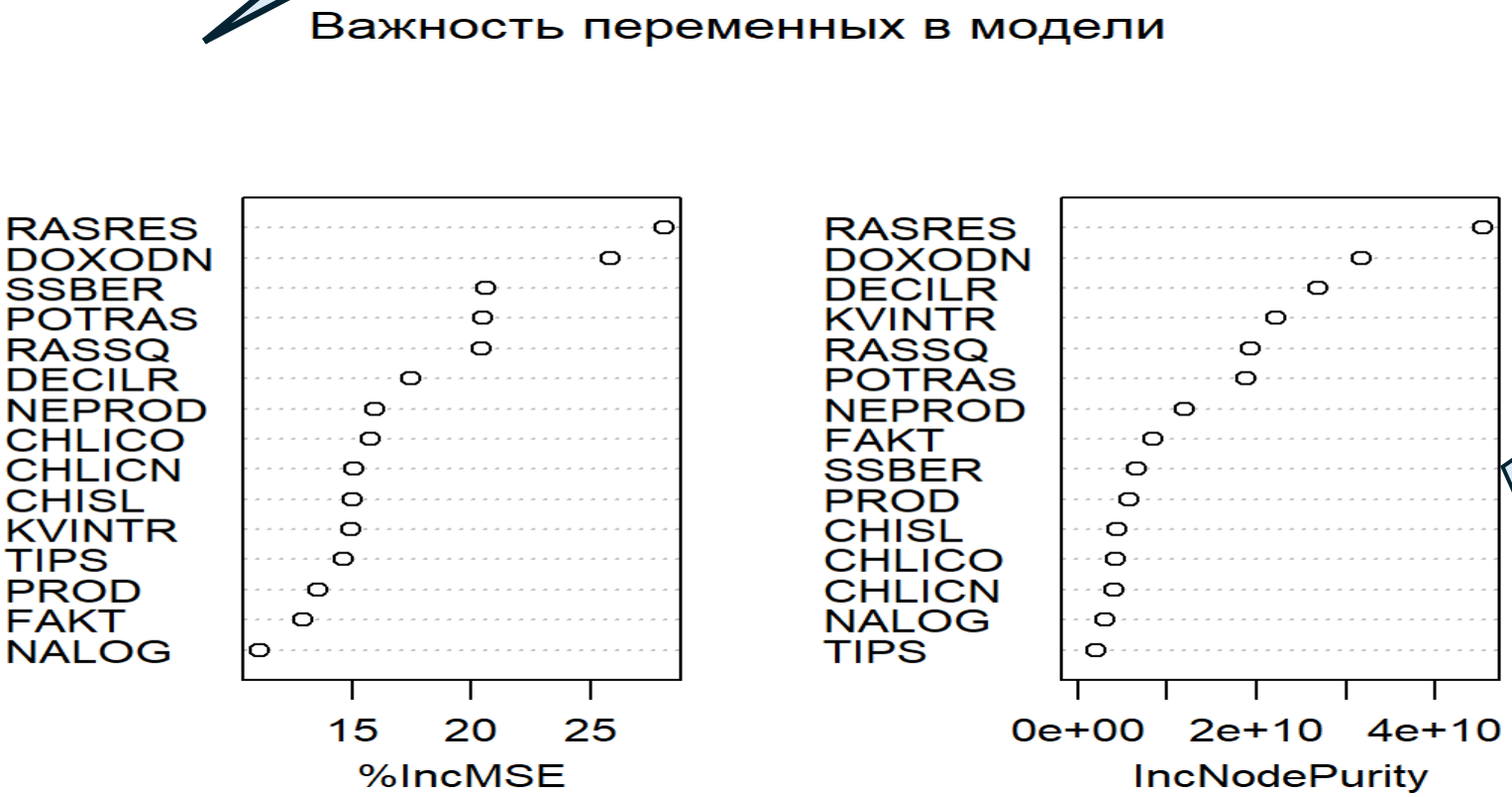
где m — количество наблюдений в тестовой выборке, y_j — истинное значение, $\hat{y}_j^{\text{ensemble}}$

Номер дерева Число узлов

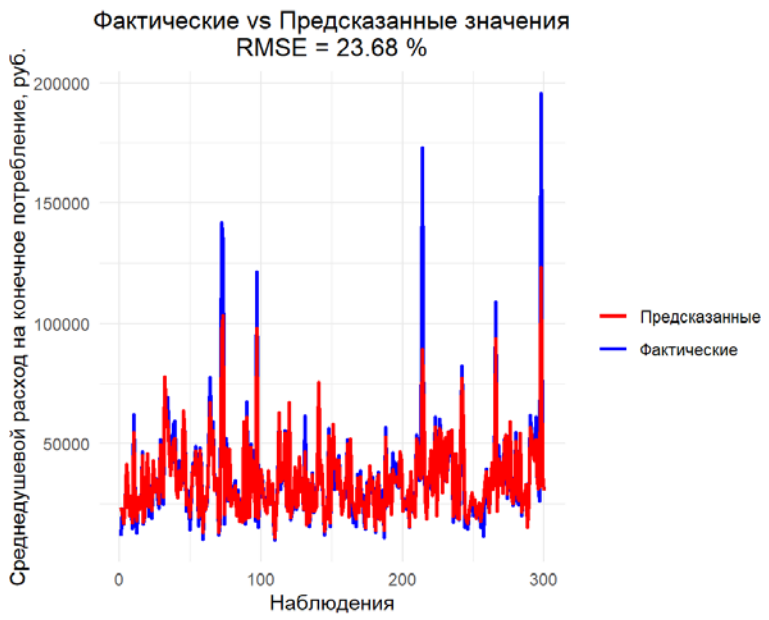
```
   tree_id number_nodes
1      183           423
2      229           425
3      139           429
4       86           433
5      175           433
6      427           433
> tail(tree_number_nodes)
    tree_id number_nodes
495     331           489
496     496           489
497      14           491
498     446           491
499     487           493
500     481           497
```

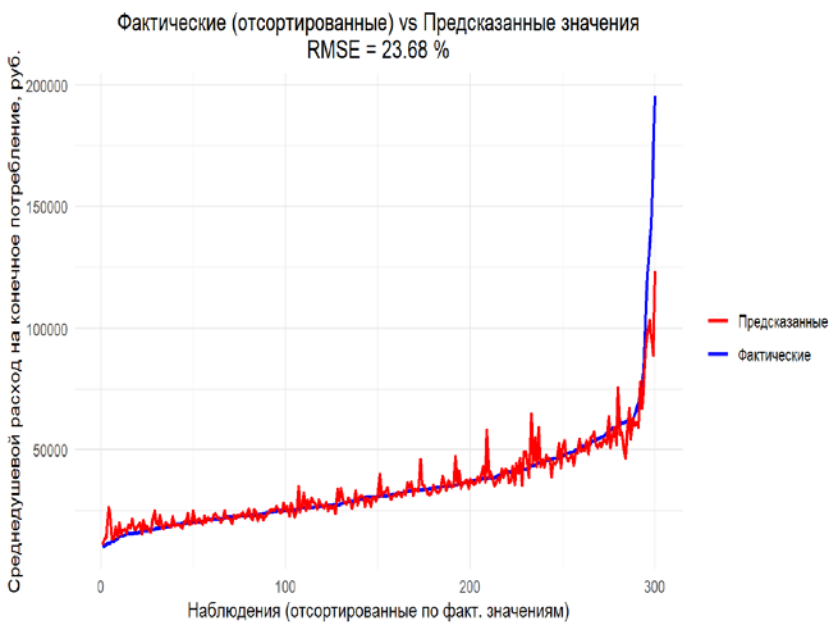
Метрики важности переменных в предсказании целевой переменной (RASQ)

% incMSE (Mean Increase in Mean Squared Error / % увеличение MSE)
Что это?
Метрика, которая показывает, насколько процентов УХУДШИТСЯ прогноз модели, если мы "испортим" эту переменную.

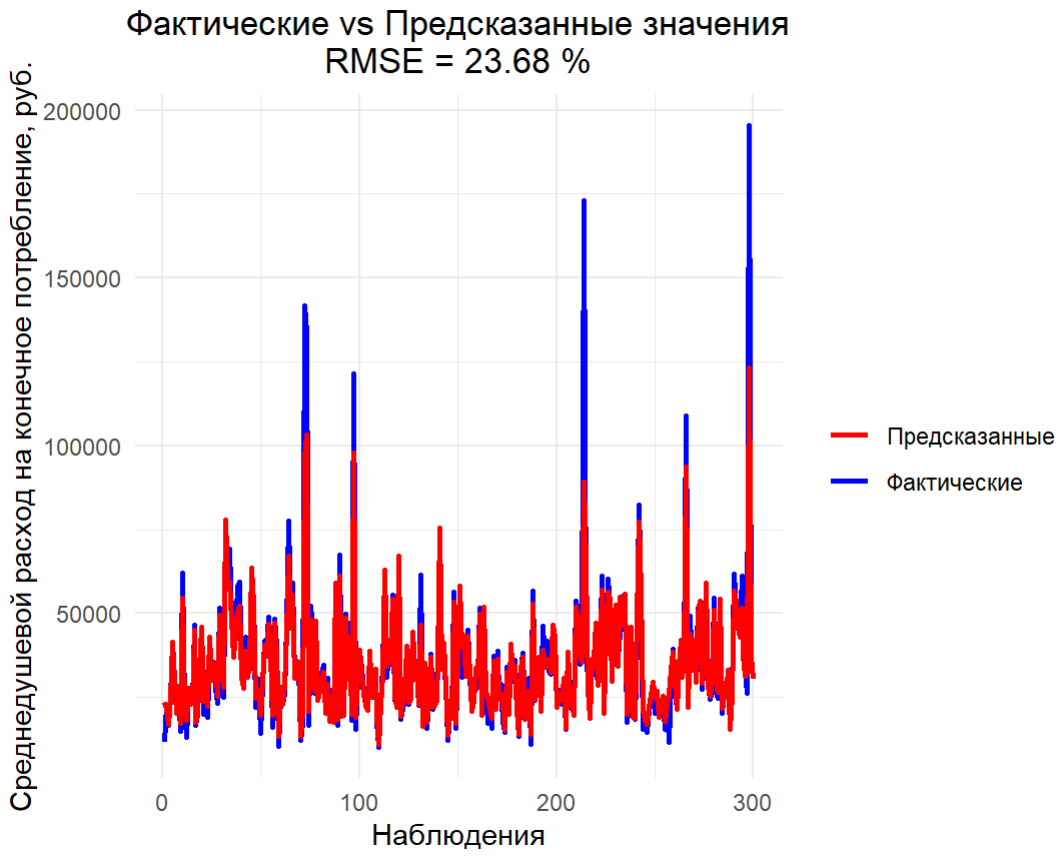


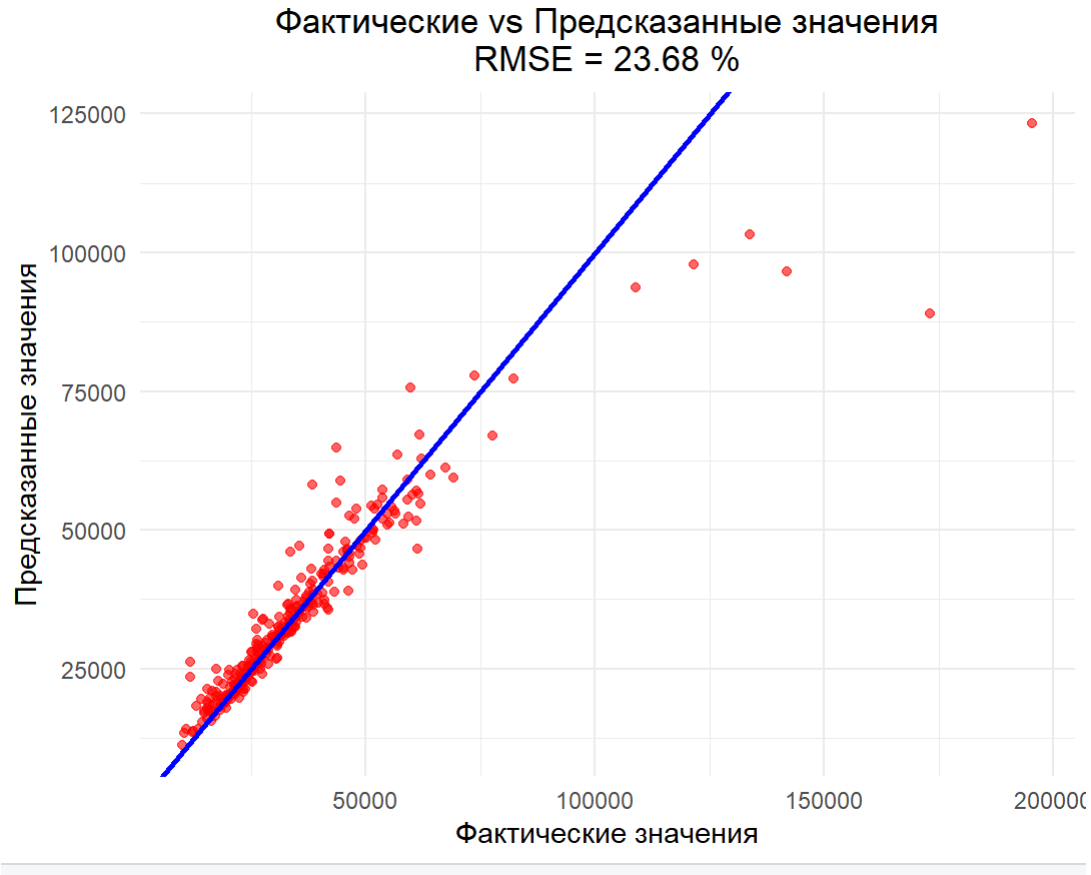
IncNodePurity (Increase in Node Purity / Прирост чистоты узла)
Что это?
Метрика, которая показывает, насколько в СРЕДНЕМ переменная повышала "чистоту" узлов при каждом использовании для разделения.





Визуализация качества модели


Фактические vs Предсказанные значения
RMSE = 23.68 %
Среднедушевой расход на конечное потребление, руб.
200000
150000
100000
50000
0
100
200
300
Наблюдения
Предсказанные
Фактические

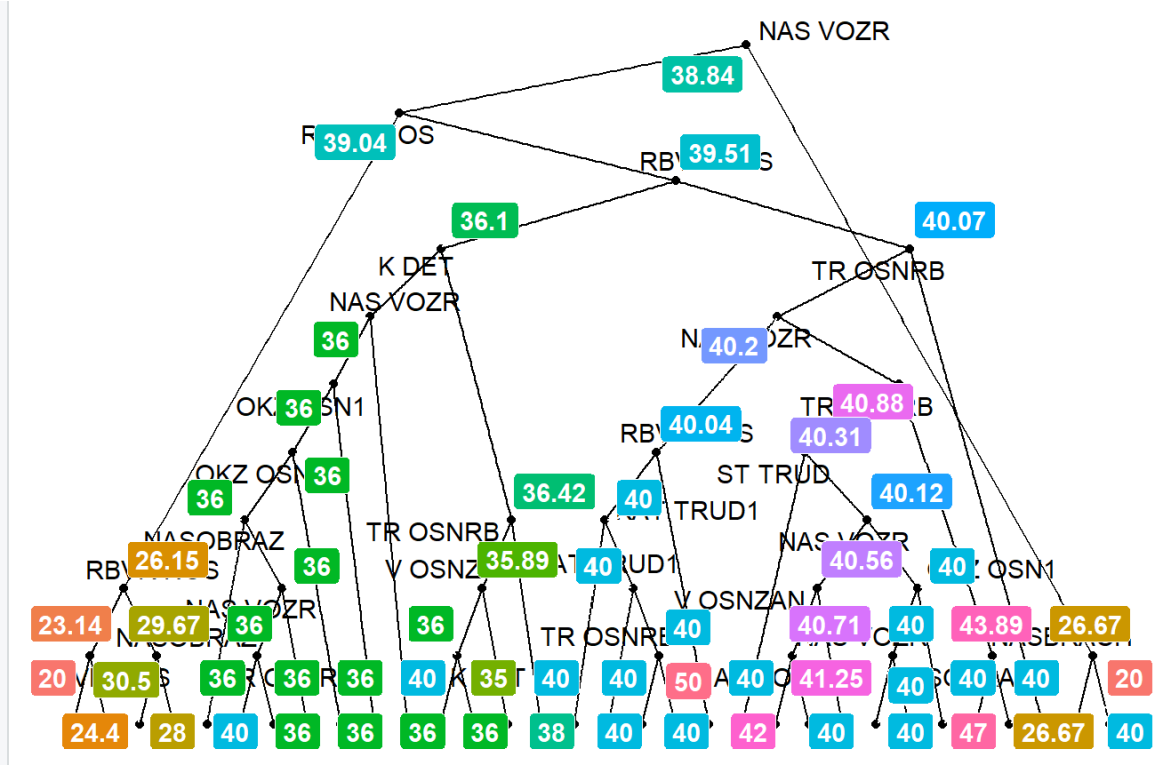

Фактические vs Предсказанные значения
RMSE = 23.68 %
Предсказанные значения
125000
100000
75000
50000
25000
50000
100000
150000
Фактические значения

Пример 3. Выявление факторов, влияющих на значение переменной RBVR_FOS – фактически отработанное время в неделю (час.). ОРС микроданные (фрагмент)

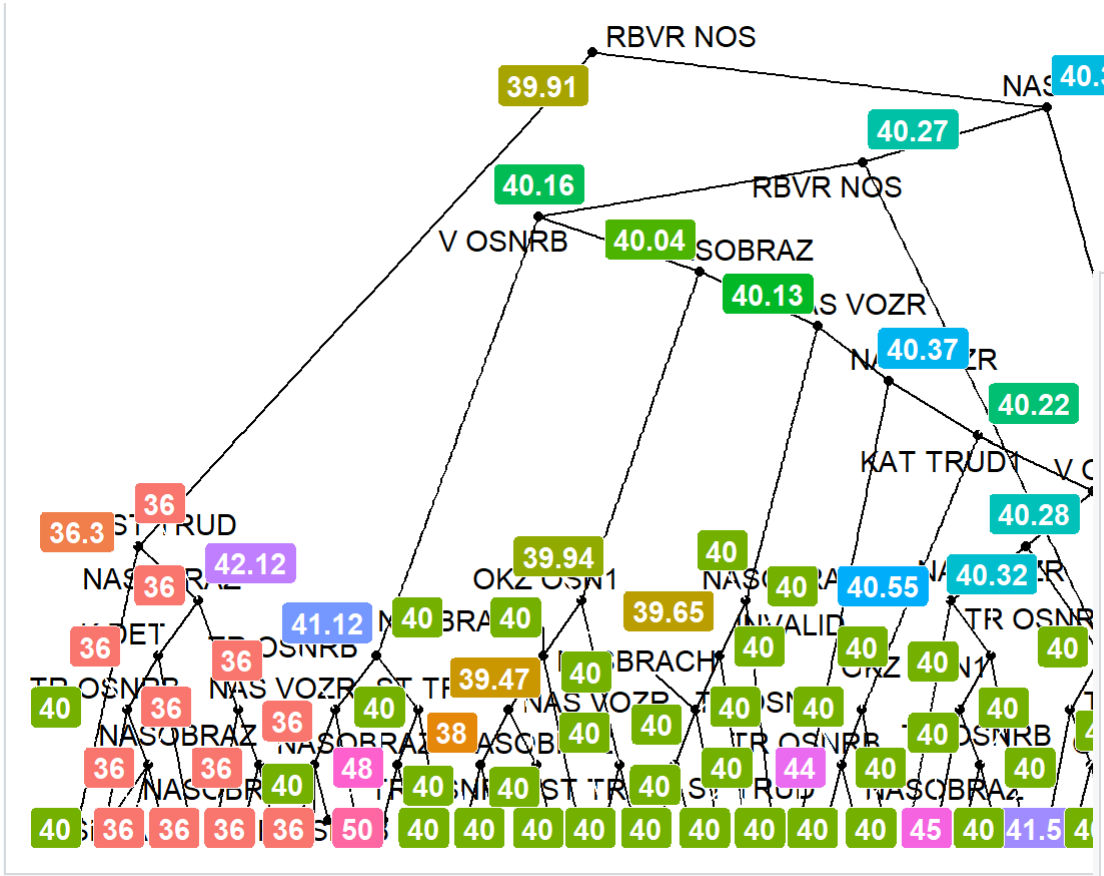
Код ОРС	
territ	Место жительства
posel	Тип населенного пункта
BB2	Число лиц в домохозяйстве
NAS_POL	Пол
NAS_VOZR	Число исполнившихся лет на 31 декабря 2019 года
NASOBRAZ	По уровню образования респондентов
INVALID	Инвалиды (всех возрастов)
V_OSNZAN	Семья с детьми по числу детей до 18 лет
KAT_TRUD1	Основная работа
V_OSNRB	Статус занятости по найму
OKZ_OSN1	По группам занятий респондентов
RBVR_FOS	Фактическая продолжительность рабочей недели в часах

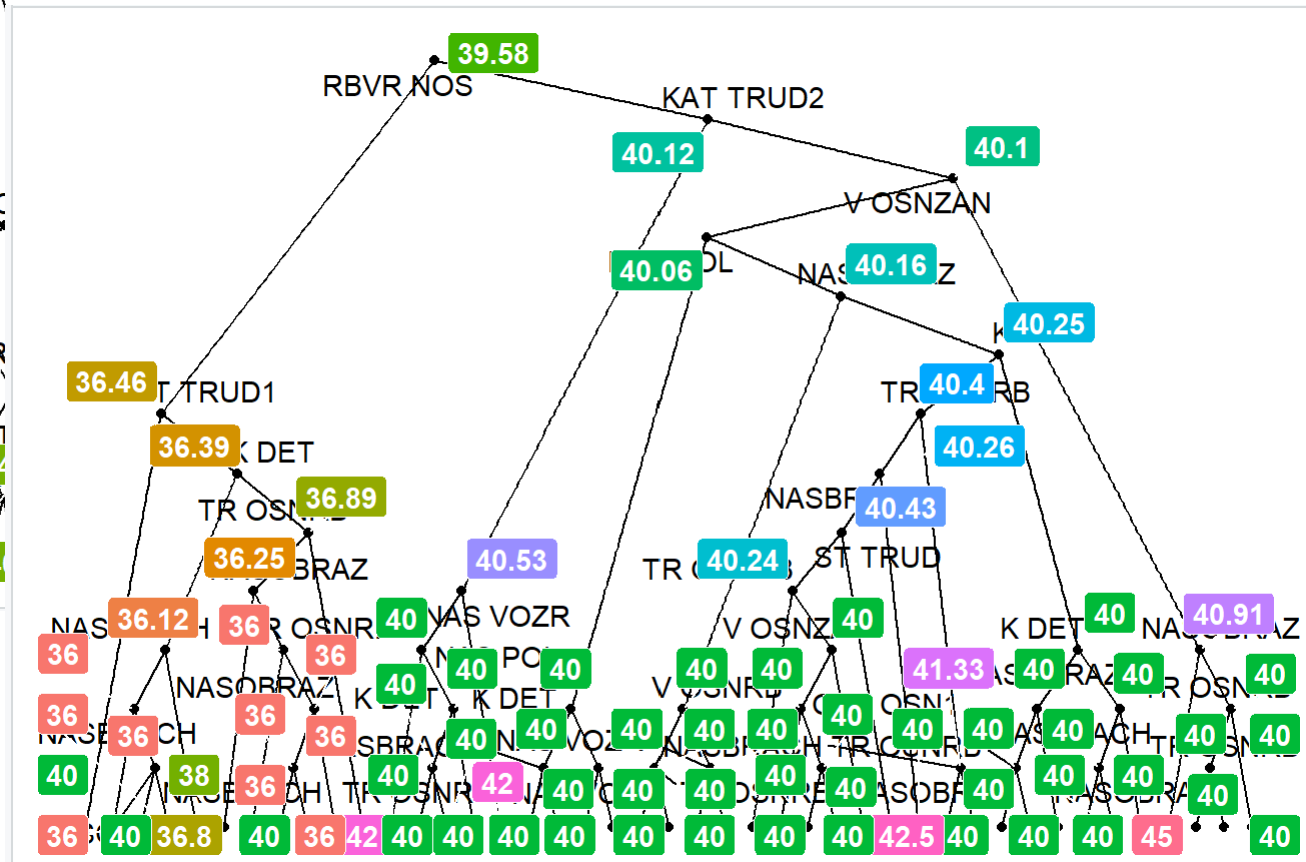
RBVR_NOS	rbvr_nos нормальная продолжительность рабочей недели на основной работе
ST_TRUD	Категория занятости по статусу
KAT_TRUD2	Категория занятости по риску
TR_OSNRB	tr_osnrb код территории основной работы
NASBRACH	nasbrach семейное положение
K_DET	k_det общее количество детей в семье, проживающих вместе с родителями
STRUKTAK	struktak - Структура населения по экономической ак
MES	Номер периода отчета
VESA_OLD	Коэффициент взвешивания ОРС первая оценка
VESA	Коэффициент взвешивания ОРС после переписи

OPC_1000 – ансамбль деревьев регрессии по переменной RBVR_FOS



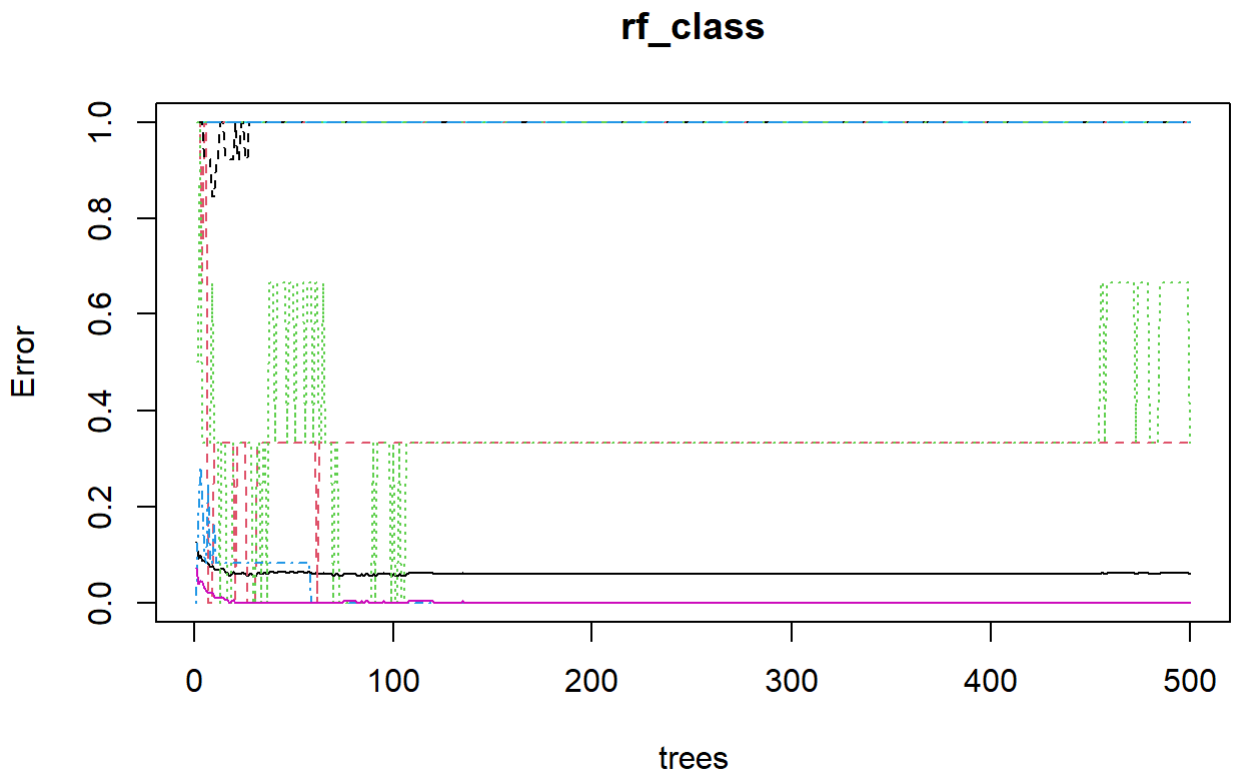
OPC_1000 (35<RBVR_F0S<=72)






Фактические (отсортированные) vs Предсказанные значения
50.0
47.5
45.0
42.5
40.0
RBVR_FOS
0
50
100
Наблюдения (отсортированные по факт. значениям)
Предсказанные
Фактические
Средняя квадратическая ошибка: 4.9 %

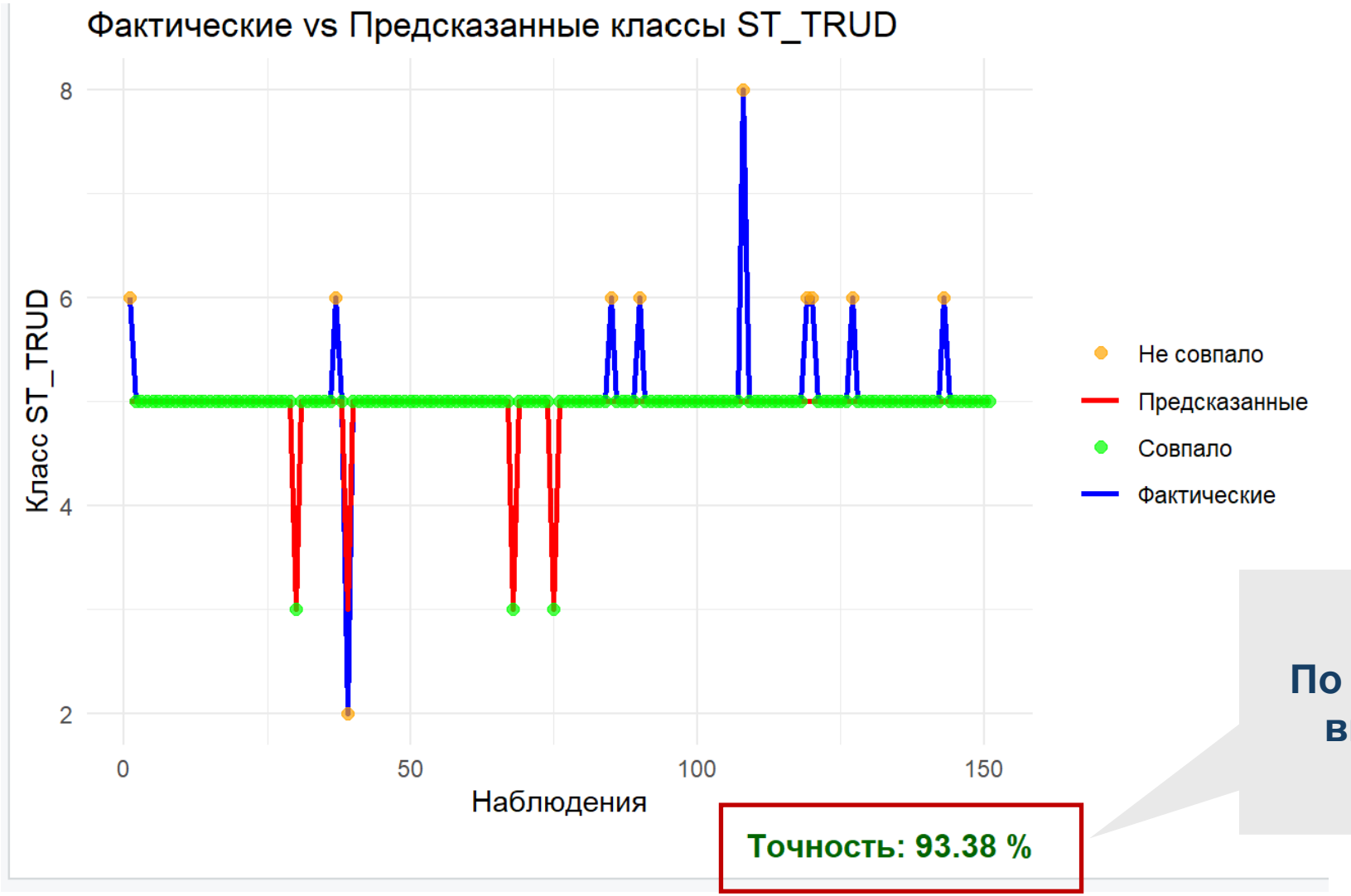
ПРИМЕР 4. Классификация по переменной ST_TRUD с применением ансамбля деревьев классификации



Черная линия: Общая ошибка классификации (OOB Error).
Красная линия: Ошибка предсказания класса "Незанятые" (0).
Зеленая линия: Ошибка предсказания класса "Занятые" (1).

- На графике показано, как меняется ошибка классификации по мере роста количества деревьев в «лесу».
- Стабильность модели: После примерно 50-100 деревьев все линии ошибок выходят на постоянный уровень. Это значит, что нам не нужно строить тысячи деревьев — модель уже стабильна и «выучила» все возможные закономерности.
- Отсутствие переобучения: Если бы ошибка на обучающих данных продолжала снижаться, а на проверочных — расти, это был бы признак переобучения. Нашем случае график показывает, что модель обобщает закономерности, а не просто заучивает данные.

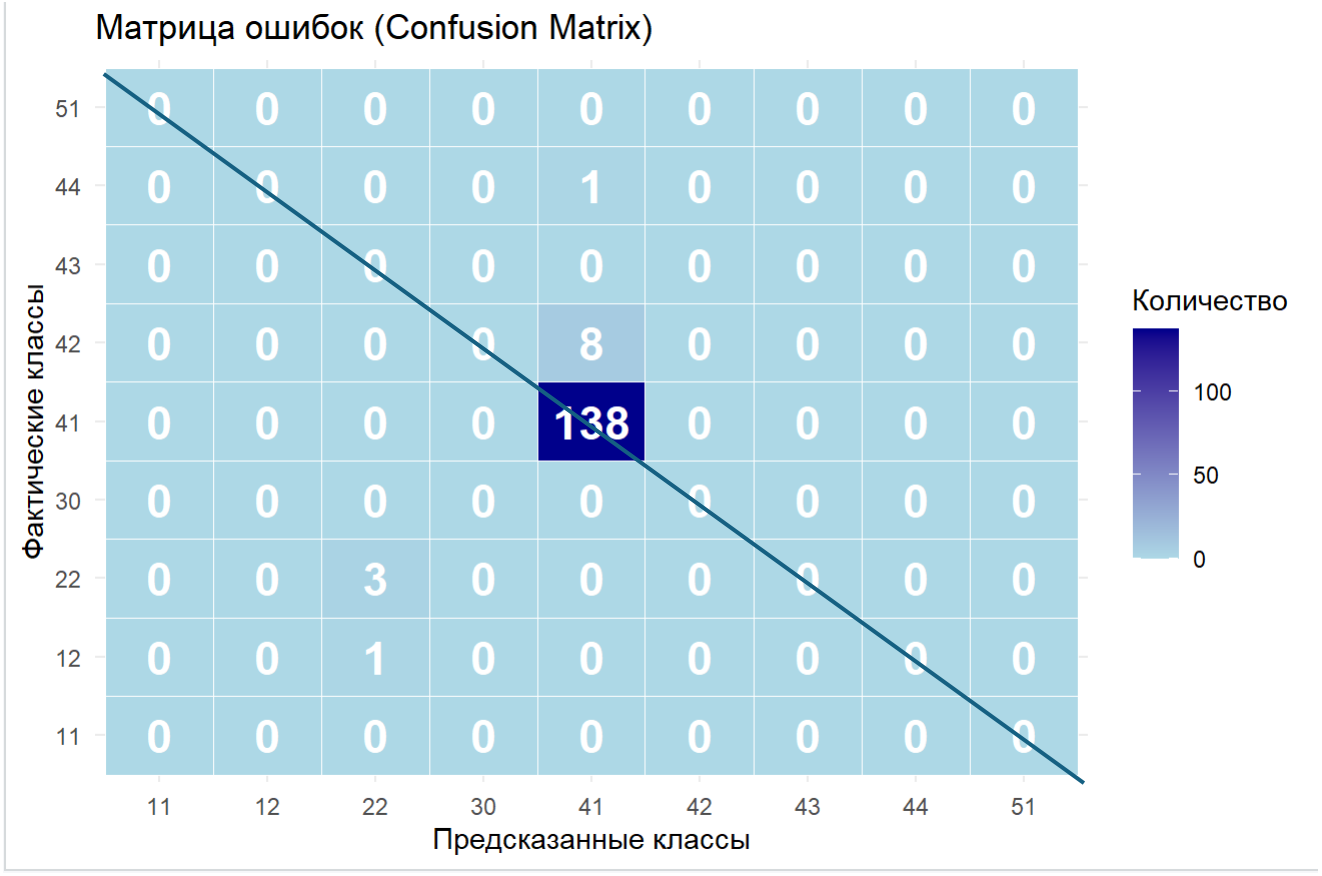
Процент совпадения фактических и предсказанных значений классификации по статусу занятости

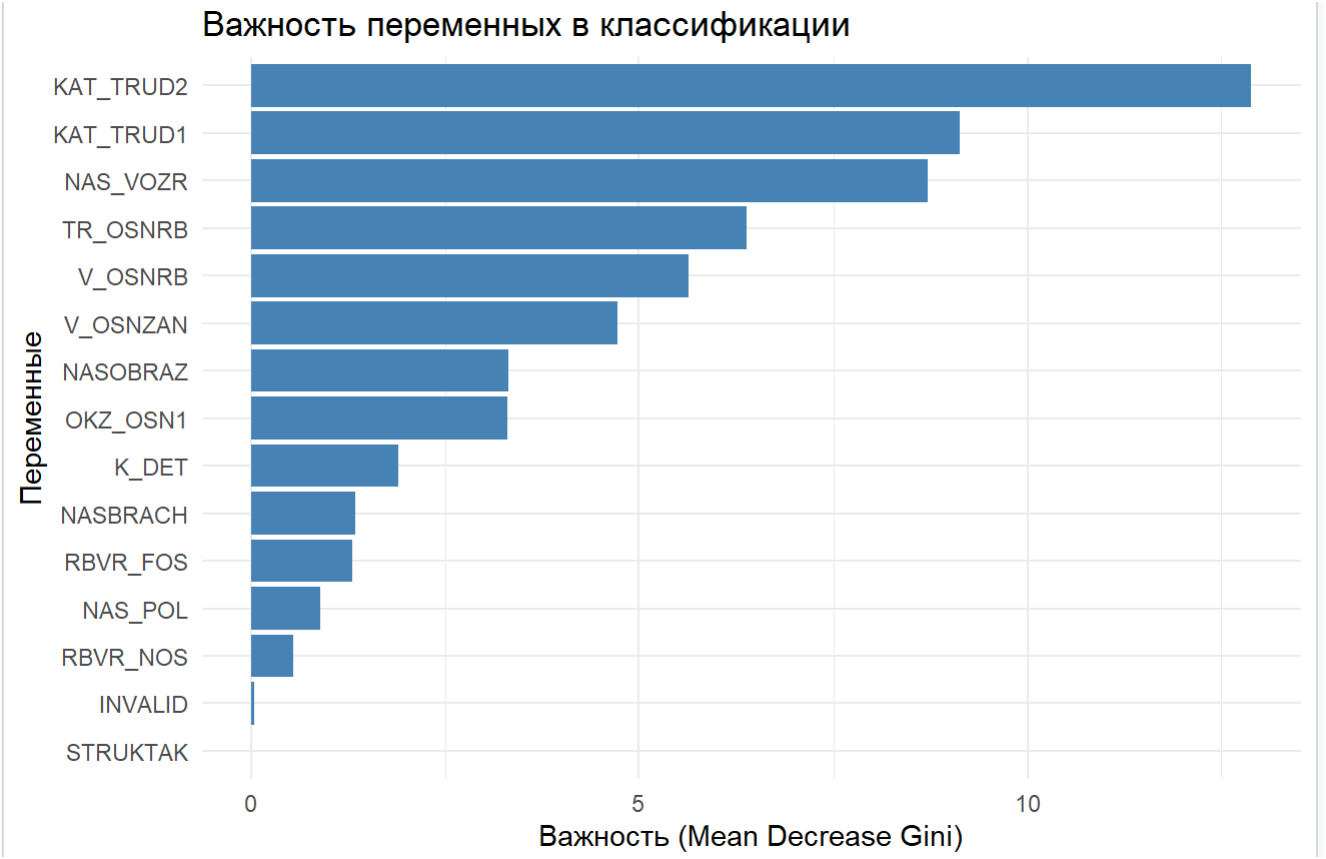


Точность: 93.38 %

По тестовой выборке

Матрица совпадений фактических и предсказанных значений классификации ST_TRUD по тестовой выборке





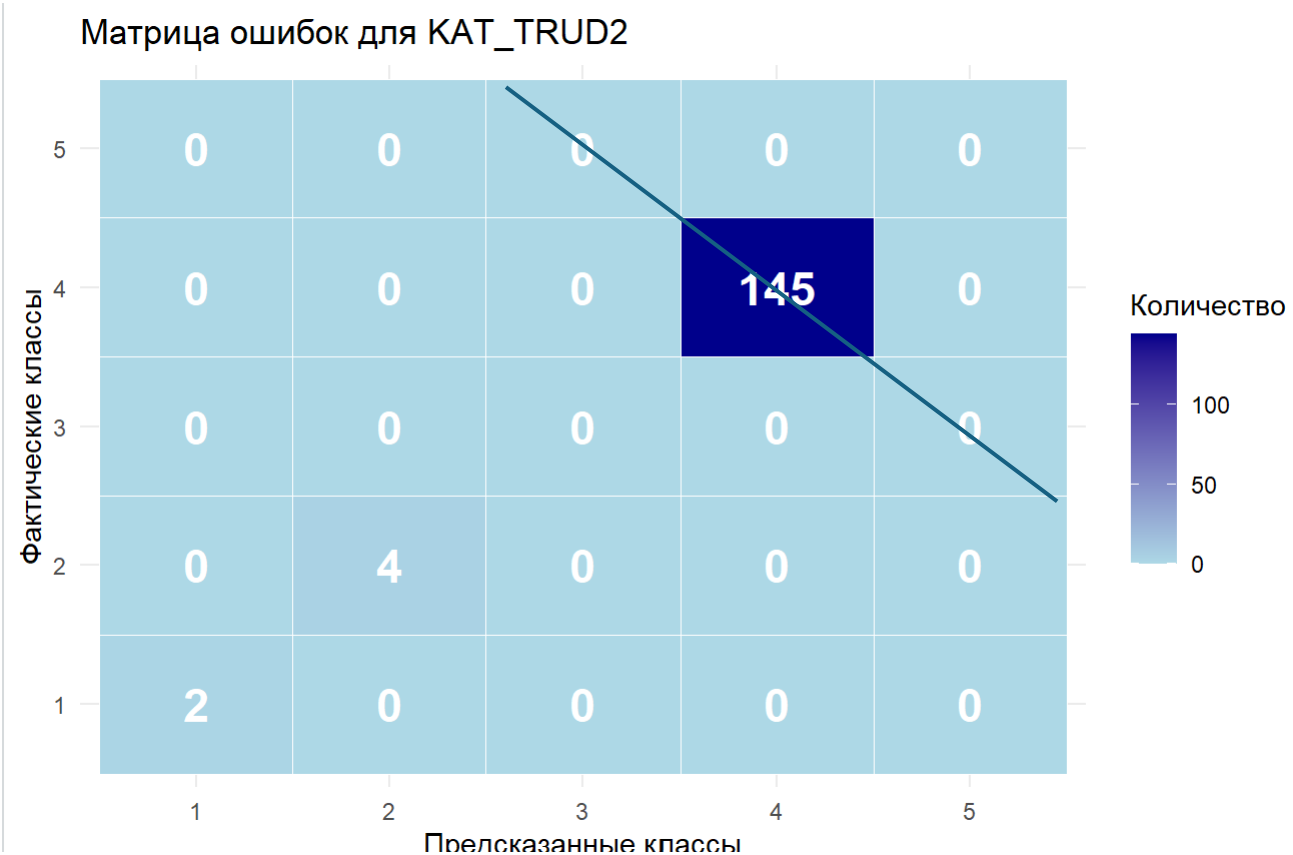
«Какое условие (например, Возраст > 35) даст нам САМОЕ ЧИСТОЕ разделение данных?»

Метрика «Mean Decrease Gini» (Среднее уменьшение неопределенности Джини) для переменной показывает:

Насколько в среднем использование этой переменной для разделения данных **снижало неопределенность (индекс Джини)** во всех деревьях леса.

Переменная с большим значением «Mean Decrease Gini» — это та переменная, которая **чаще и лучше всего** помогала алгоритму создавать

KUT_TRUD2


Матрица ошибок для KAT_TRUD2
Фактические классы
Предсказанные классы
Количество
100
50
0
1
2
3
4
5
0
0
0
0
0
0
0
0
145
0
0
0
0
0
0
0
0
0
0
4
0
0
0
2
0
0
0
0

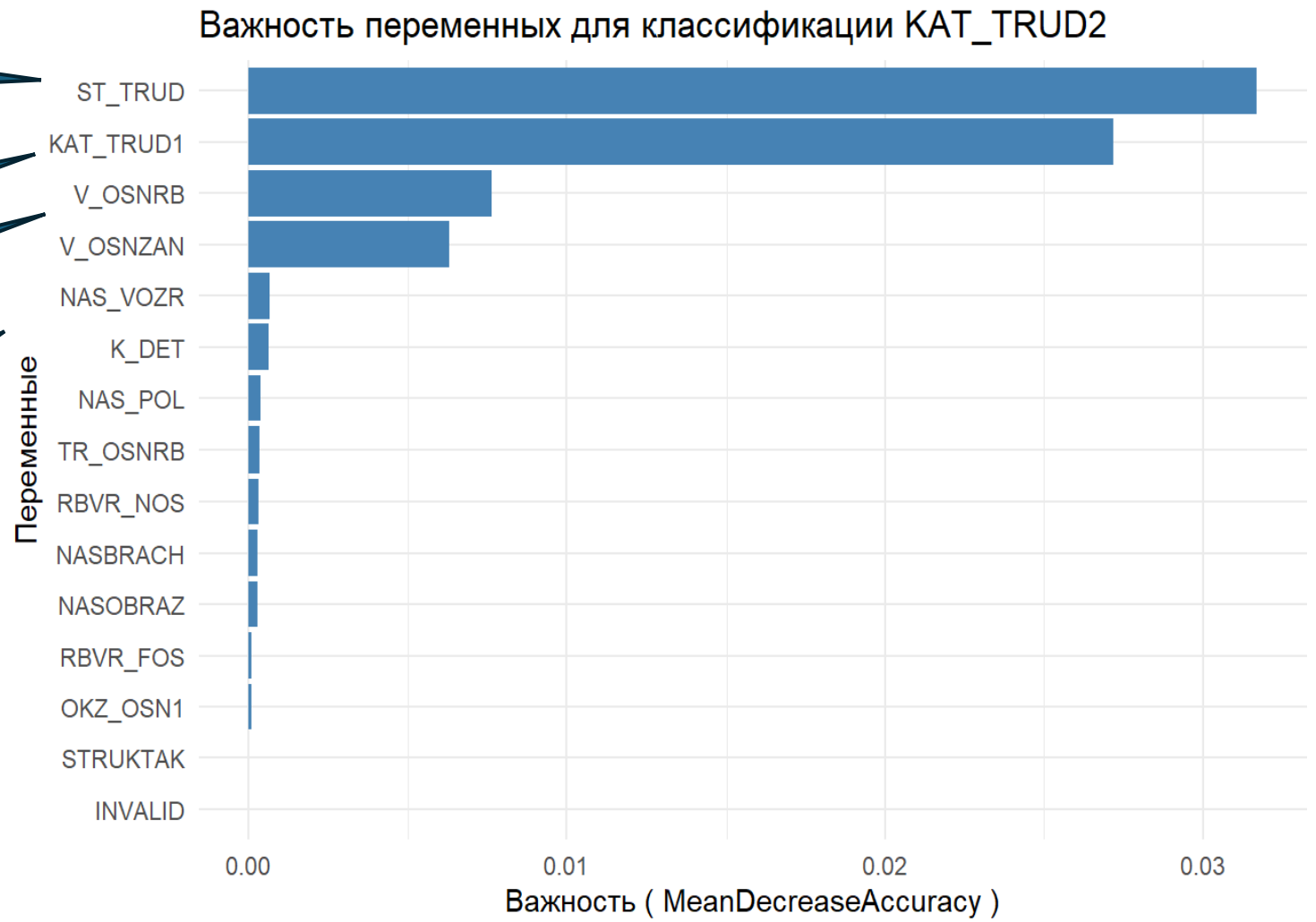
Наиболее важные переменные в классификации

Статус занятости

Основная работа

Статус занятости по найму

Возраст



Применение в официальной статистике стран СНГ

01

Анализ демографических данных

Классификация населения по социально-экономическим группам, прогнозирование миграционных потоков и оценка демографических трендов на основе множественных факторов.

02

Экономическое моделирование

Регрессионные деревья для прогнозирования макроэкономических показателей, анализа факторов роста ВВП и оценки влияния политических решений на экономику.

03

Социальная статистика

Выявление факторов бедности, классификация домохозяйств по уровню благосостояния, анализ доступности социальных услуг в различных регионах.

04

Энергетическая безопасность

Оценка устойчивости энергетических систем с помощью ансамблей деревьев (случайный лес) в России и Казахстане, прогнозирование потребления энергоресурсов.

Перспектива развития: интеграция методов машинного обучения в официальную статистику открывает новые возможности для улучшения качества прогнозов, принятия обоснованных решений и формирования эффективной государственной политики.

Практические кейсы по странам СНГ

Россия

Росстат применяет деревья решений для анализа промышленного производства, классификации предприятий по уровню инновационности и прогнозирования региональных экономических показателей. Случайный лес используется для оценки устойчивости энергосистем.

Казахстан

Статагентство РК внедрило модели машинного обучения для анализа факторов экономического роста регионов, классификации домохозяйств по благосостоянию и прогнозирования демографических процессов в условиях урбанизации.

Беларусь

Белстат использует регрессионные деревья для анализа сельскохозяйственного производства, прогнозирования урожайности и классификации предприятий АПК по эффективности. Модели помогают оптимизировать аграрную политику.

Узбекистан

Госкомстат применяет деревья классификации для анализа рынка труда, выявления факторов безработицы и оценки эффективности программ профессионального образования. Модели учитывают региональную специфику.

Общий тренд: статистические службы СНГ активно внедряют методы машинного обучения для повышения качества аналитики и поддержки принятия решений на государственном уровне.