

ML-ПРОЕКТ

House Prices - Advanced Regression Techniques

Команда DS 2N_Siberia
2022





ЭТАП 1: ВВОДНЫЕ ДАННЫЕ

ЦЕЛЬ:

Решение задачи House Prices - Advanced Regression Techniques*

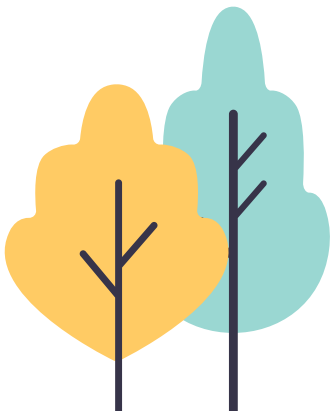
ЗАДАЧИ:

1. Отработка навыков работы с признаками
2. Отработка навыков применения алгоритмов машинного обучения (ML, регрессия)
3. Отработка навыка командного взаимодействия

КОМАНДА ПРОЕКТА:

- КАЗАНЦЕВ Егор
- КОНЧЕВ Александр
- ЧЕРЕПАНОВА Ирина
- БОГОВЕЕВ Дмитрий

* - <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>



ЭТАП 1: ВВОДНЫЕ ДАННЫЕ

ПРОЕКТ: House Prices - Advanced Regression Techniques:

Проект представляет собой определение (предсказание) наиболее вероятной стоимости домовладений в Эймсе, штат Айова. Основой для предсказания стоимости являются 79 независимых переменных (признаков), описывающих различные аспекты жилых домов



ДОРОЖНАЯ КАРТА ПРОЕКТА





ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)



1460

СТРОК

79

СТОЛБЦОВ



115 340

ПАРАМЕТРОВ

NaN

Выборки содержат
пропуски данных

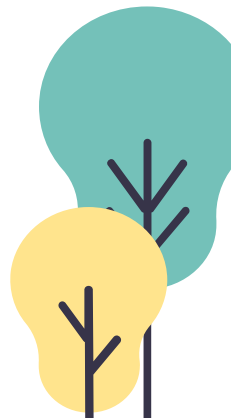


Abc123

Содержат категориальные и
числовые признаки

ВЫБРОСЫ

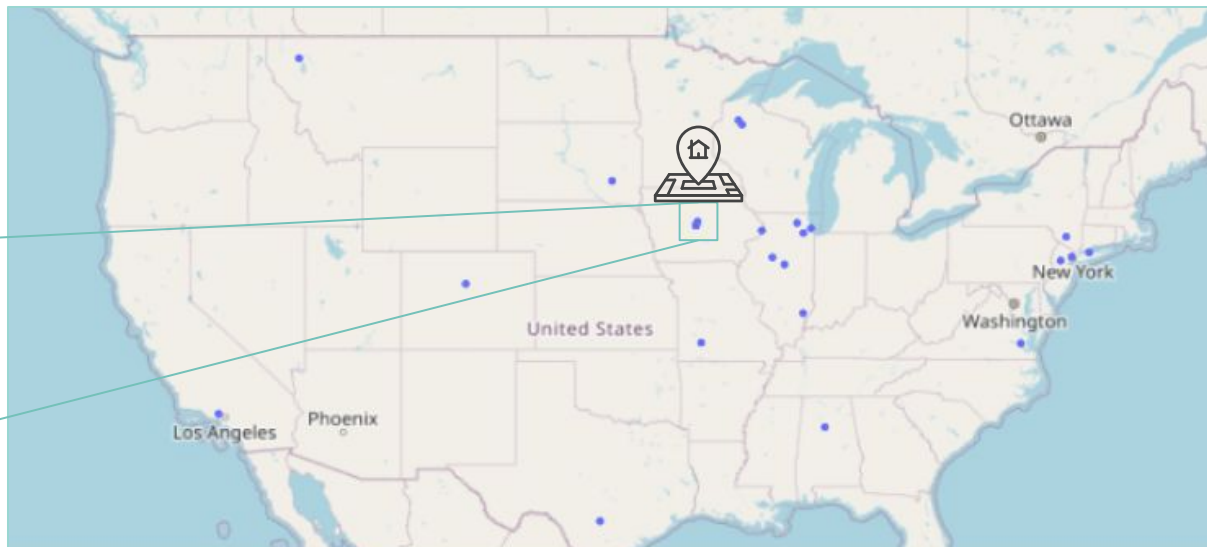
Данные не однородны



ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

1

Изучение
месторасположения домов



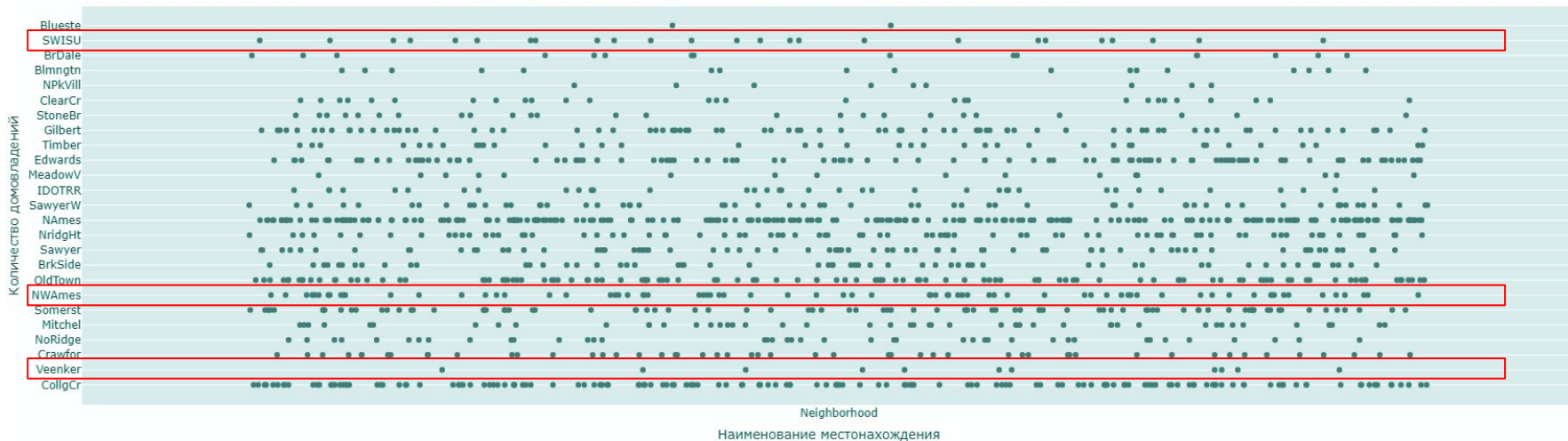
ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

1

Изучение
месторасположения домов



Распределение домовладений из обучающей выборки по месторасположению



ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

1

Изучение
месторасположения домов



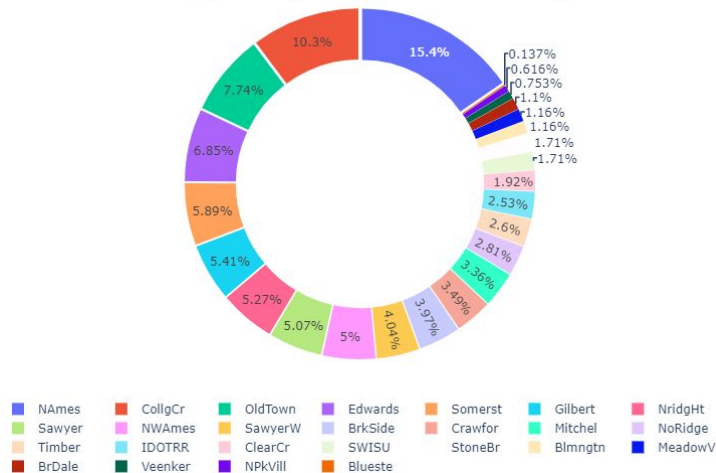
ВЫВОДЫ:

Только 3 из 25 мест расположения домов в выборке относятся к штату Айова (17,8% всех домов).
Корреляция территориального признака со стоимостью - 0,21.

РЕШЕНИЕ:

Сохранить признак в виду малозначимости

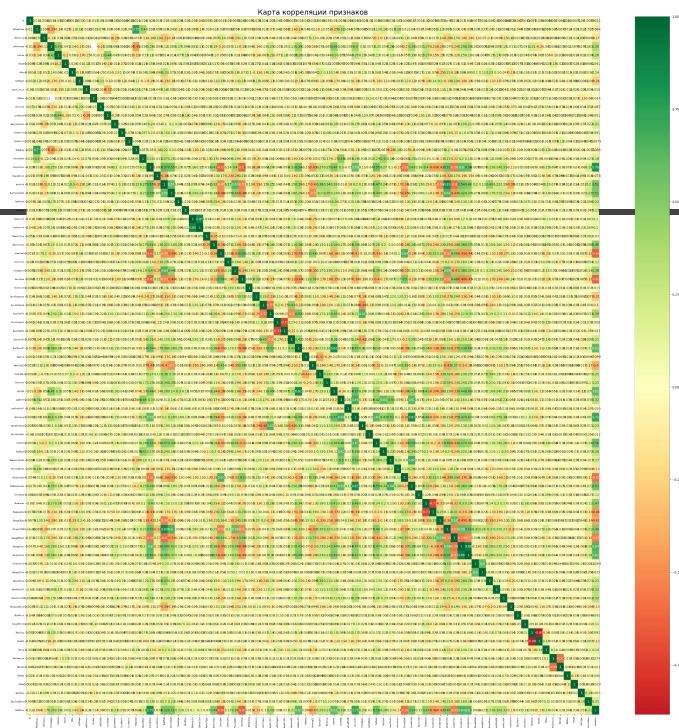
Распределение домовладений по местонахождению



ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

2

Изучение корреляции
признаков и стоимости
дома



ВЫВОДЫ:

Признаки относительно сбалансированы.

Наибольшая корреляция признаков, положительная

- OverallQual (общее состояние отделки дома)
- GrLivArea (площадь помещений над землей)

Отрицательная

- ExterQual (оценка качества материала снаружи дома)
- BsmtQual (Высота потолка в подвале)

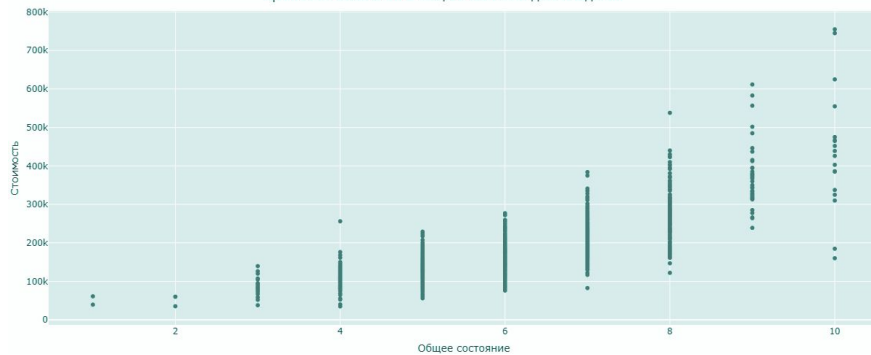
ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

2

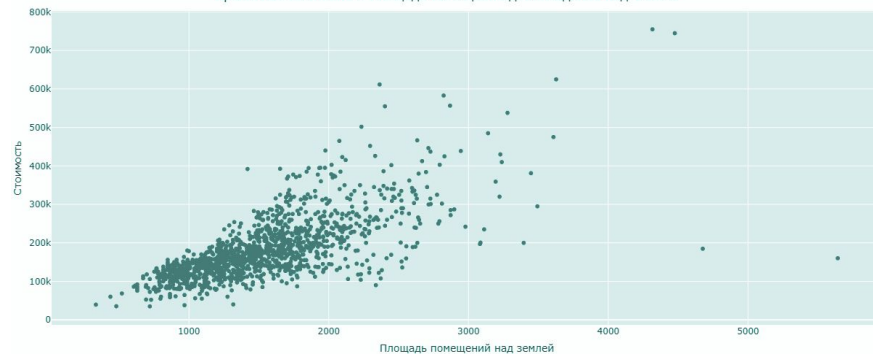
Изучение корреляции
признаков и стоимости
дома



Сравнение стоимости и общего состояния домовладения



Сравнение стоимости и площади помещений домовладения над землей



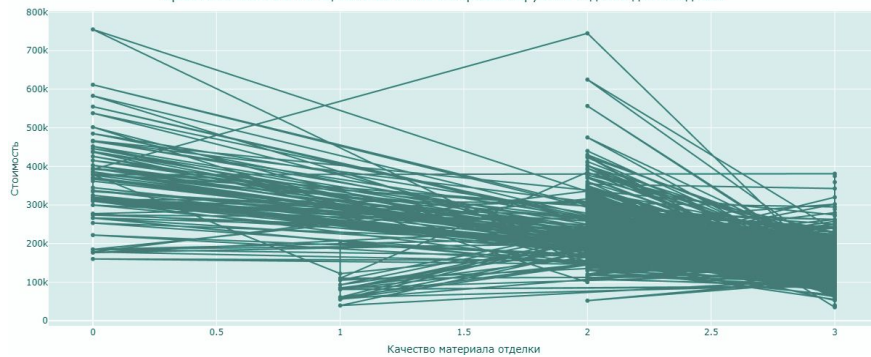
ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)

2

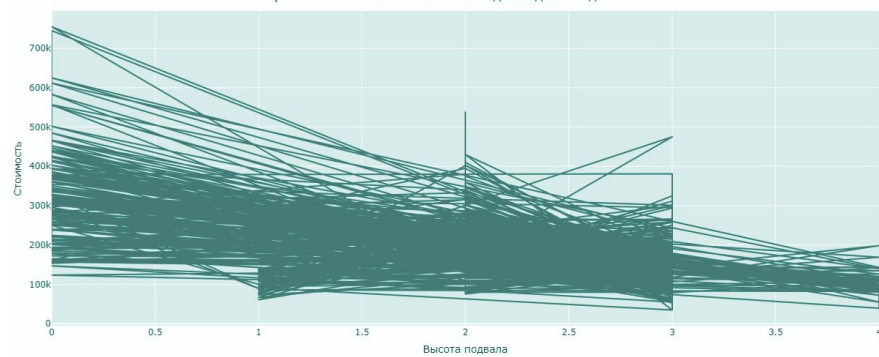
Изучение корреляции
признаков и стоимости
дома



Сравнение стоимости и оценкой качества материала наружной отделки домовладения



Сравнение стоимости и высотой подвала домовладения



ЭТАП 2: ИЗУЧЕНИЕ И ПОДГОТОВКА ПРИЗНАКОВ (ПЕРЕМЕННЫХ)



ВЫВОДЫ:

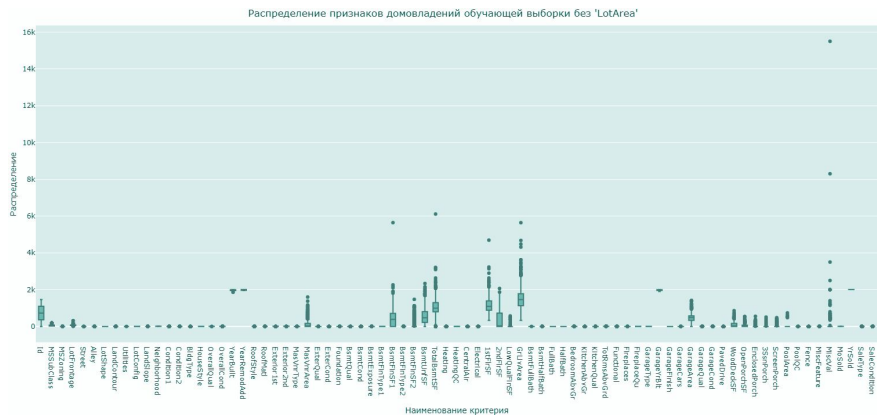
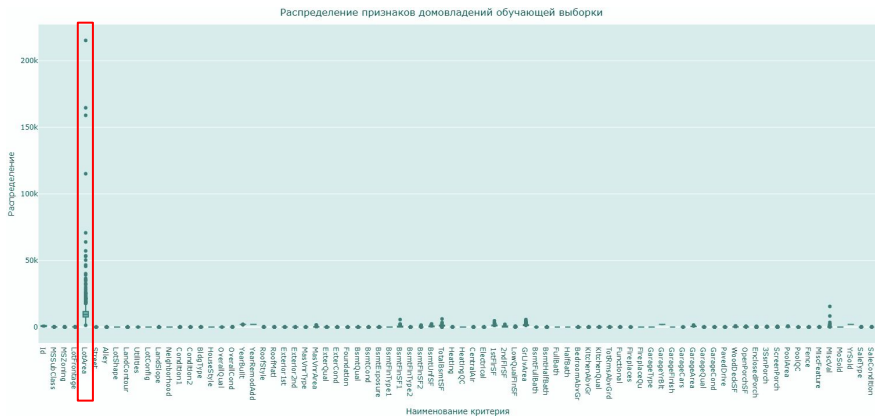
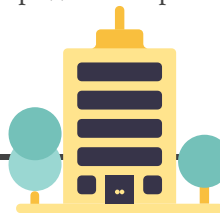
Признак **LotArea** (размер участка), имеет большие выбросы, при корреляции 0,26.

РЕШЕНИЕ:

В виду малозначимости и большой неоднородности - удалить признак.

3

Однородность признаков



ЭТАП 3: ПОДГОТОВКА ВЫБОРОК (ДАТАСЕТОВ)

4

Предобработка выборок



**Выборки
готовы для
применения
алгоритмов**



ВЫВОДЫ:

- перекодированы категориальные признаки
- выборки разделены в пропорции **75/25** на обучающую, тестовую (контрольную), целевая выборка осталась без разделения
- пропуски значений (**NaN**) заменены средними по выборке
- определены алгоритмы машинного обучения (**Random forest, Линейная регрессия, ElasticNet, Градиентный бустинг**)
- определены метрики качества алгоритмов (**RMSE, R2**)

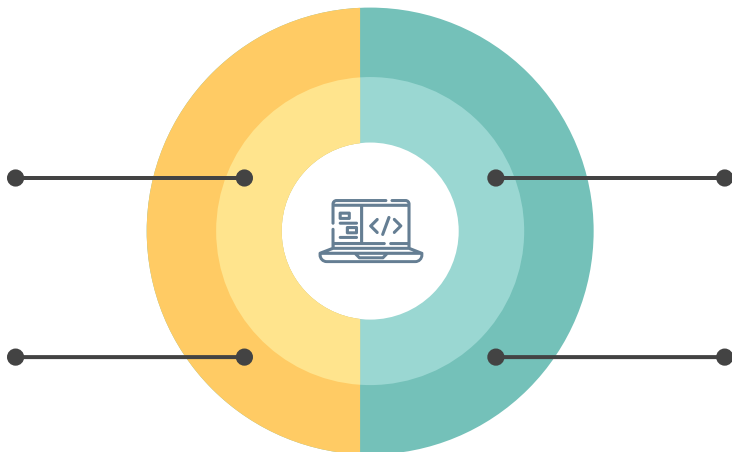
ЭТАП 4: МАШИННОЕ ОБУЧЕНИЕ

Алгоритм Random Forest

КАЗАНЦЕВ Егор

Алгоритм Линейной регрессии

ЧЕРЕПАНОВА Ирина



Алгоритм ElasticNet

БОГОВЕЕВ Дмитрий

Алгоритм Градиентный бустинг

КОЧНЕВ Александр

ЭТАП 5: МЕТРИКА R2

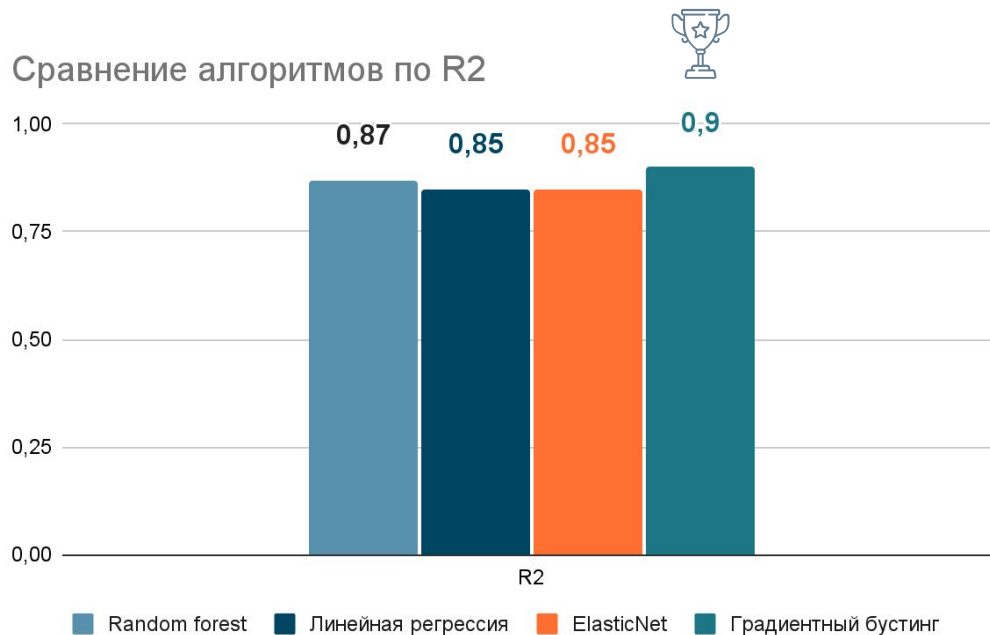
1
Random Forest

2
Линейная регрессия

3
ElasticNet

4
Градиентный бустинг

Сравнение алгоритмов по R2



ЭТАП 5: МЕТРИКА RMSE

1

Random Forest

2

Линейная регрессия

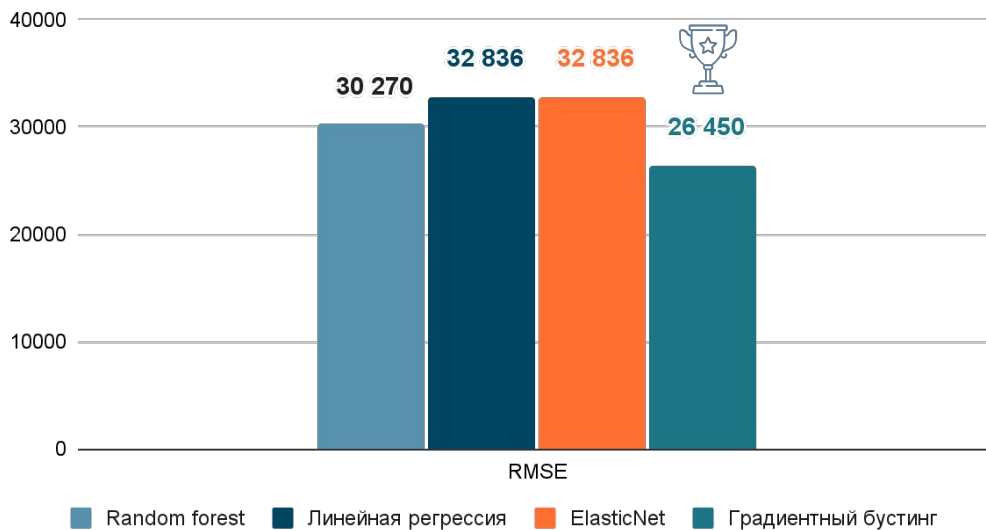
3

ElasticNet

4






Градиентный бустинг

Сравнение алгоритмов по RMSE



ЭТАП 5: РЕЗУЛЬТАТ ПРОЕКТА

РЕЗУЛЬТАТ (Градиентный бустинг) РЕШЕНИЯ ЗАДАЧИ НА KAGGLE:

1644	DS 2N_Siberia	   	0.13655	1	1s
 Your First Entry! Welcome to the leaderboard!					





СПАСИБО!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.

