

Проект: House Price Predictions

Работу выполнили:
Юрий Смолий, Елена Ефимова, Мария Королева

Задача — предсказание цены продажи домов

Кратко о датасете:

80 признаков (класс здания, форма участка, стиль дома и т.п.)

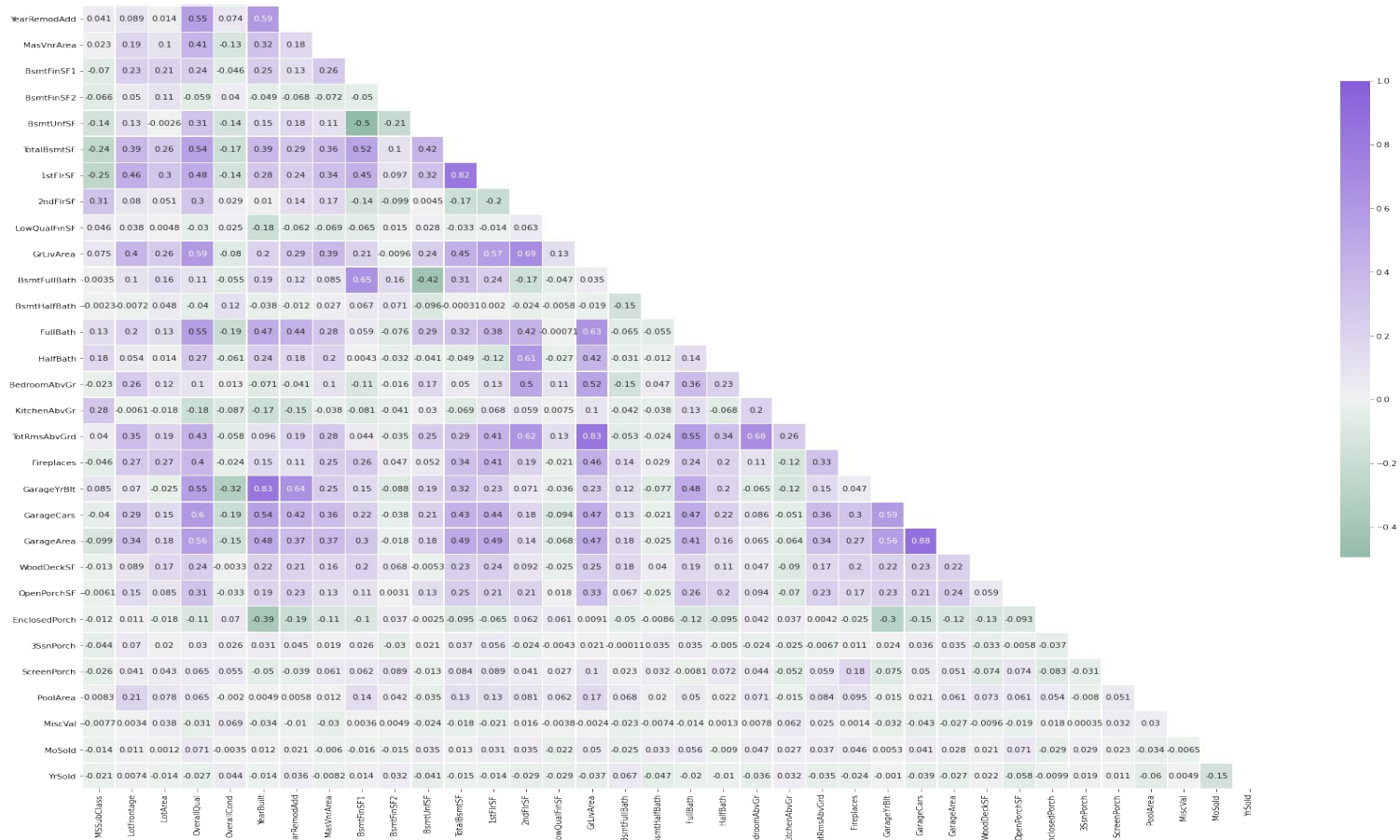
train — 1460 строк

test — 1459 строк

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	...	ScreenPorch	PoolArea	PoolQC	Fence	Mi
0	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	
1	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	...	0	0	NaN	NaN	
2	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	0	NaN	NaN	
3	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	0	NaN	NaN	
4	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	...	0	0	NaN	NaN	
...
1455	60	RL	62.0	7917	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	
1456	20	RL	85.0	13175	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	MnPrv	
1457	70	RL	66.0	9042	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	GdPrv	
1458	20	RL	68.0	9717	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	
1459	20	RL	75.0	9937	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	

1460 rows x 79 columns

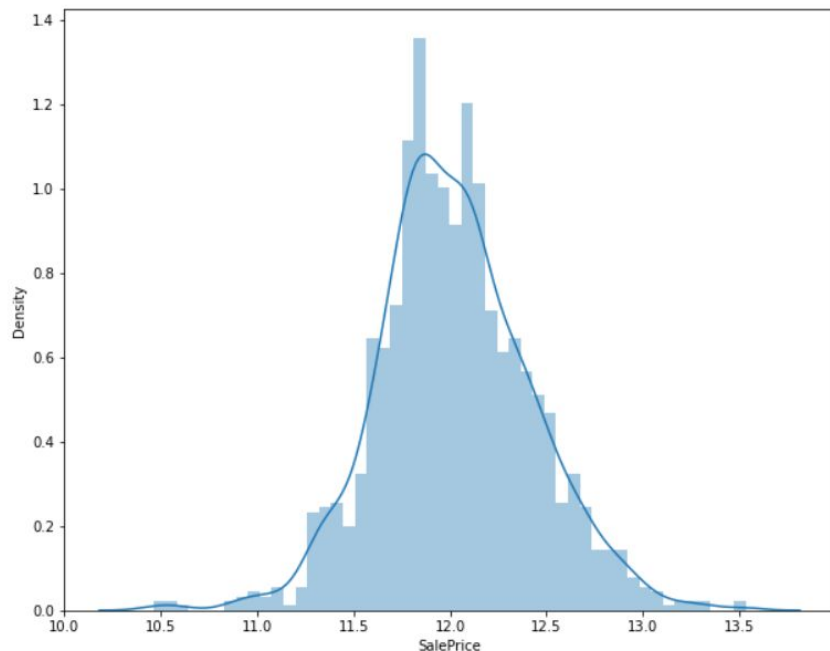
Матрица корреляции признаков



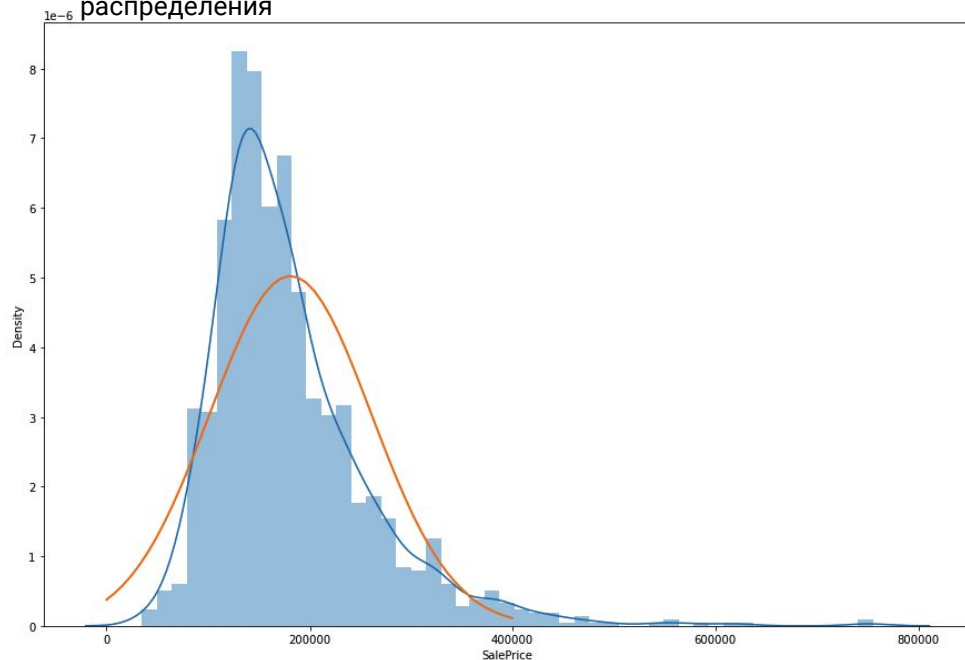
Исследование целевой переменной

Распределение целевой переменной напоминает нормальное. Проверив гипотезу о нормальности распределения тестом Колмогорова-Смирнова, гипотезу отвергаем ($p_value=0.000$)

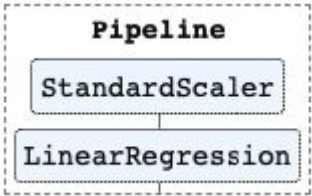
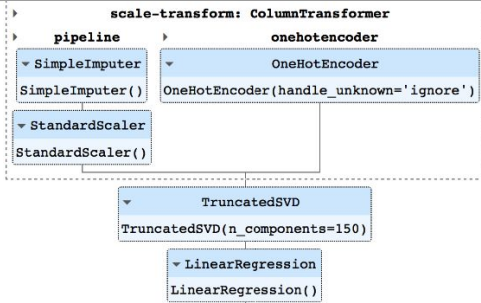
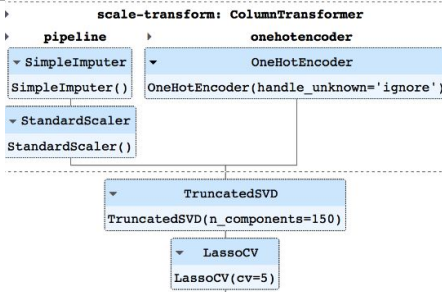
Логарифмированный таргет



Целевая переменная без обработки с графиком нормального распределения



Модели (линейные):

Описание	Схема	Score
Обработка только числовых признаков с удалением категориальных без учета корреляции. Использована стандартизация данных StandardScaler	 <pre> graph TD Pipeline[Pipeline] --> StandardScaler[StandardScaler] StandardScaler --> LinearRegression[LinearRegression] </pre>	0.44892
Линейная регрессия с уменьшением размерности с помощью усеченного SVD(TruncatedSVD), использована стандартизация StandardScaler, категориальные признаки обработаны OneHotEncoder(пропуски игнорируются). Пропуски вещественных признаков заменены на среднее.	 <pre> graph TD Pipeline[Pipeline] --> SimpleImputer[SimpleImputer] SimpleImputer --> StandardScaler[StandardScaler] StandardScaler --> OneHotEncoder[OneHotEncoder] OneHotEncoder --> TruncatedSVD[TruncatedSVD] TruncatedSVD --> LinearRegression[LinearRegression] </pre>	0.15132
Метод LassoCV(количество участков для кросс-валидации (параметр cv) равен 5)с уменьшением размерности с помощью усеченного SVD(TruncatedSVD), использована стандартизация StandardScaler, категориальные признаки обработаны OneHotEncoder(пропуски игнорируются). Пропуски вещественных признаков заменены на среднее.	 <pre> graph TD Pipeline[Pipeline] --> SimpleImputer[SimpleImputer] SimpleImputer --> StandardScaler[StandardScaler] StandardScaler --> OneHotEncoder[OneHotEncoder] OneHotEncoder --> TruncatedSVD[TruncatedSVD] TruncatedSVD --> LassoCV[LassoCV] </pre>	0.14958

Модели (логистическая регрессия и случайный лес):

Описание	Схема	Score
<p>Логистическая регрессия</p> <p>Логистическая регрессия с уменьшением размерности с помощью усеченного SVD (TruncatedSVD), использована стандартизация StandardScaler, категориальные признаки обработаны OneHotEncoder (пропуски игнорируются). Пропуски вещественных признаков заменены на среднее</p>	<pre> graph TD subgraph scale_transform [scale-transform: ColumnTransformer] subgraph pipeline SI[SimpleImputer] --> SS[StandardScaler] end subgraph onehotencoder OHE[OneHotEncoder] end end pipeline --> TSVD[TruncatedSVD] onehotencoder --> TSVD TSVD --> LR[LogisticRegression] </pre>	0.23017
<p>RandomForest</p> <p>Использован случайный лес. Категориальные признаки обработаны OneHotEncoder (пропуски игнорируются). Пропуски вещественных признаков заменены на медиана. Использована стандартизация StandardScaler. Подбор гиперпараметров для случайного леса по GridSearch</p>	<pre> graph TD subgraph scale_transform [scale-transform: ColumnTransformer] subgraph pipeline SI[SimpleImputer] --> SS[StandardScaler] end subgraph onehotencoder OHE[OneHotEncoder] end end pipeline --> LSCV[LinearSVC] onehotencoder --> LSCV LSCV --> RF[RandomForestRegressor] </pre>	0.15132

Модели (градиентный бустинг):

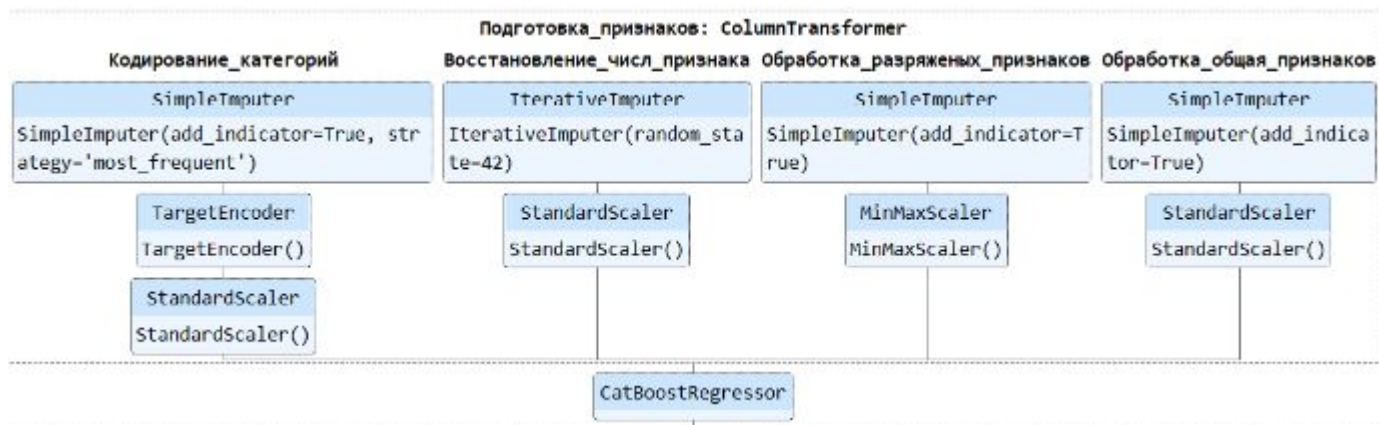
Описание	Схема	Score
<p>Градиентный бустинг №1</p> <p>Использовался градиентный Бустинг. Сложная обработка признаков. Добавлены полиномиальные признаки. квантильная трансформация, робастная стандартизация, логарифмированный таргет</p>	<pre> graph TD subgraph Pipeline subgraph Кодирование_категорий SI1[SimpleImputer SimpleImputer(add_indicator=True, strategy='most_frequent')] QE[QuantileEncoder QuantileEncoder()] SI1 --> QE end subgraph Подготовка_признаков_ColumnTransformer subgraph Восстановление_числ_признака II1[IterativeImputer IterativeImputer(random_sta te=42)] QT1[QuantileTransformer QuantileTransformer()] II1 --> QT1 end subgraph Обработка_разряженных_признаков SI2[SimpleImputer SimpleImputer(add_indicator=T rue, strategy='median')] QT2[QuantileTransformer QuantileTransformer()] SI2 --> QT2 end subgraph Обработка_общая_признаков SI3[SimpleImputer SimpleImputer(add_indicator=T rue, strategy='median')] QT3[QuantileTransformer QuantileTransformer()] SI3 --> QT3 end subgraph Обработка_общая_признаков_2 SI4[SimpleImputer SimpleImputer(add_indicator=T rue, strategy='median')] QT4[QuantileTransformer QuantileTransformer()] SI4 --> QT4 end end QE --> PF[PolynomialFeatures PolynomialFeatures()] QT1 --> PF QT2 --> PF QT3 --> PF QT4 --> PF PF --> RS[RobustScaler RobustScaler(quantile_range=(3, 97))] RS --> SKB[SelectKBest SelectKBest(k=1000)] SKB --> GBR[GradientBoostingRegressor GradientBoostingRegressor()] end </pre>	0.13807
<p>Градиентный бустинг №2</p> <p>Обычный градиентный бустинг на стандартизованных признаках. Без доп обогащения признаками</p>	<pre> graph TD subgraph Pipeline subgraph Кодирование_категорий SI1[SimpleImputer SimpleImputer(add_indicator=True, strategy='most_frequent')] TE[TargetEncoder TargetEncoder()] SS1[StandardScaler StandardScaler()] SI1 --> TE TE --> SS1 end subgraph Подготовка_признаков_ColumnTransformer subgraph Восстановление_числ_признака II1[IterativeImputer IterativeImputer(random_sta te=42)] SS2[StandardScaler StandardScaler()] II1 --> SS2 end subgraph Обработка_разряженных_признаков SI2[SimpleImputer SimpleImputer(add_indicator=T rue)] MMS[MinMaxScaler MinMaxScaler()] SI2 --> MMS end subgraph Обработка_общая_признаков_1 SI3[SimpleImputer SimpleImputer(add_indica tor=True)] SS3[StandardScaler StandardScaler()] SI3 --> SS3 end subgraph Обработка_общая_признаков_2 SI4[SimpleImputer SimpleImputer(add_indica tor=True)] SS4[StandardScaler StandardScaler()] SI4 --> SS4 end end SS1 --> GBR[GradientBoostingRegressor GradientBoostingRegressor()] SS2 --> GBR MMS --> GBR SS3 --> GBR SS4 --> GBR end </pre>	0.13062

Модели (CatBoost):

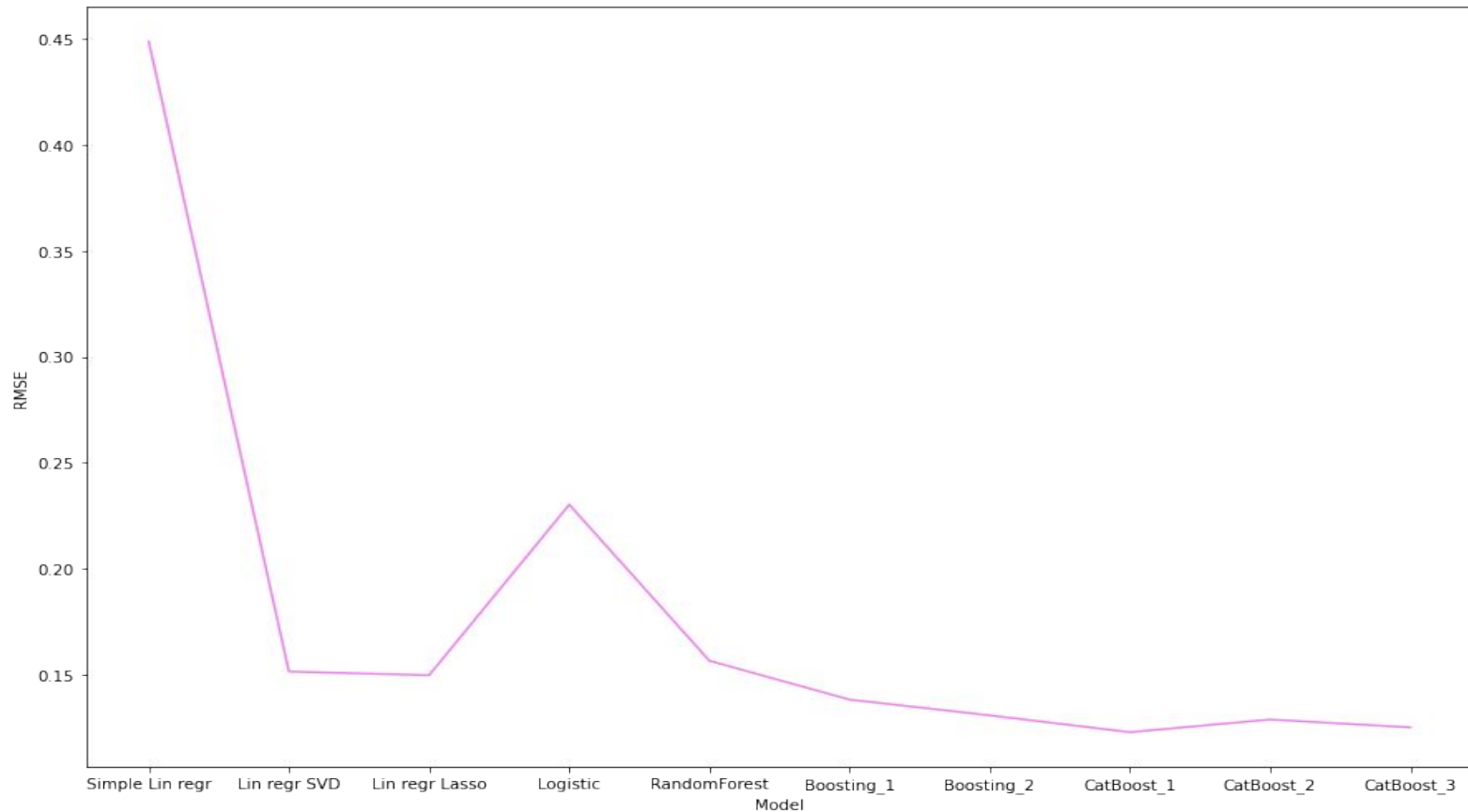
Описание	Схема	Score
<p>CatBoost №1</p> <p>Использован градиентный спуск CatBoost. Категориальные признаки обработаны OneHotEncoder (пропуски игнорируются). Пропуски вещественных признаков заменены на медиану. Использована стандартизация StandardScaler. Удалены сильно коррелирующие признаки</p>	<pre> graph TD subgraph "scale-transform: ColumnTransformer pipeline" SI[SimpleImputer] --> SS[StandardScaler] SI --> OHE[OneHotEncoder] end SS --> SFM[feature_selection: SelectFromModel] OHE --> SFM subgraph SFM L[LinearSVC] end SFM --> CB[CatBoostRegressor] </pre>	0.12269
<p>CatBoost №2</p> <p>Сложная обработка признаков. Добавлены полиномиальные признаки, квантильная трансформация, робастная стандартизация, логарифмированный таргет, катбуст на гиперпараметрах "по-умолчанию"</p>	<pre> graph TD subgraph "Подготовка_признаков: ColumnTransformer" subgraph "Кодирование_категорий" SI1[SimpleImputer SimpleImputer(add_indicator=True, strategy='most_frequent')] --> QE1[QuantileEncoder() QuantileEncoder()] end subgraph "Восстановление_числ_признака" II[IterativeImputer IterativeImputer(random_state=42)] --> QT2[QuantileTransformer() QuantileTransformer()] end subgraph "Обработка_разряженных_признаков" SI2[SimpleImputer SimpleImputer(add_indicator=True, strategy='median')] --> QT3[QuantileTransformer() QuantileTransformer()] end subgraph "Обработка_общая_признаков" SI3[SimpleImputer SimpleImputer(add_indicator=True, strategy='median')] --> QT4[QuantileTransformer() QuantileTransformer()] end end QE1 --> PF[PolynomialFeatures PolynomialFeatures()] QT2 --> PF QT3 --> PF QT4 --> PF PF --> RS[RobustScaler RobustScaler(quantile_range=(3, 97))] RS --> SKB[SelectKBest SelectKBest(k=1000)] SKB --> CB[CatBoostRegressor] </pre>	0.12867

Модели (CatBoost продолжение):

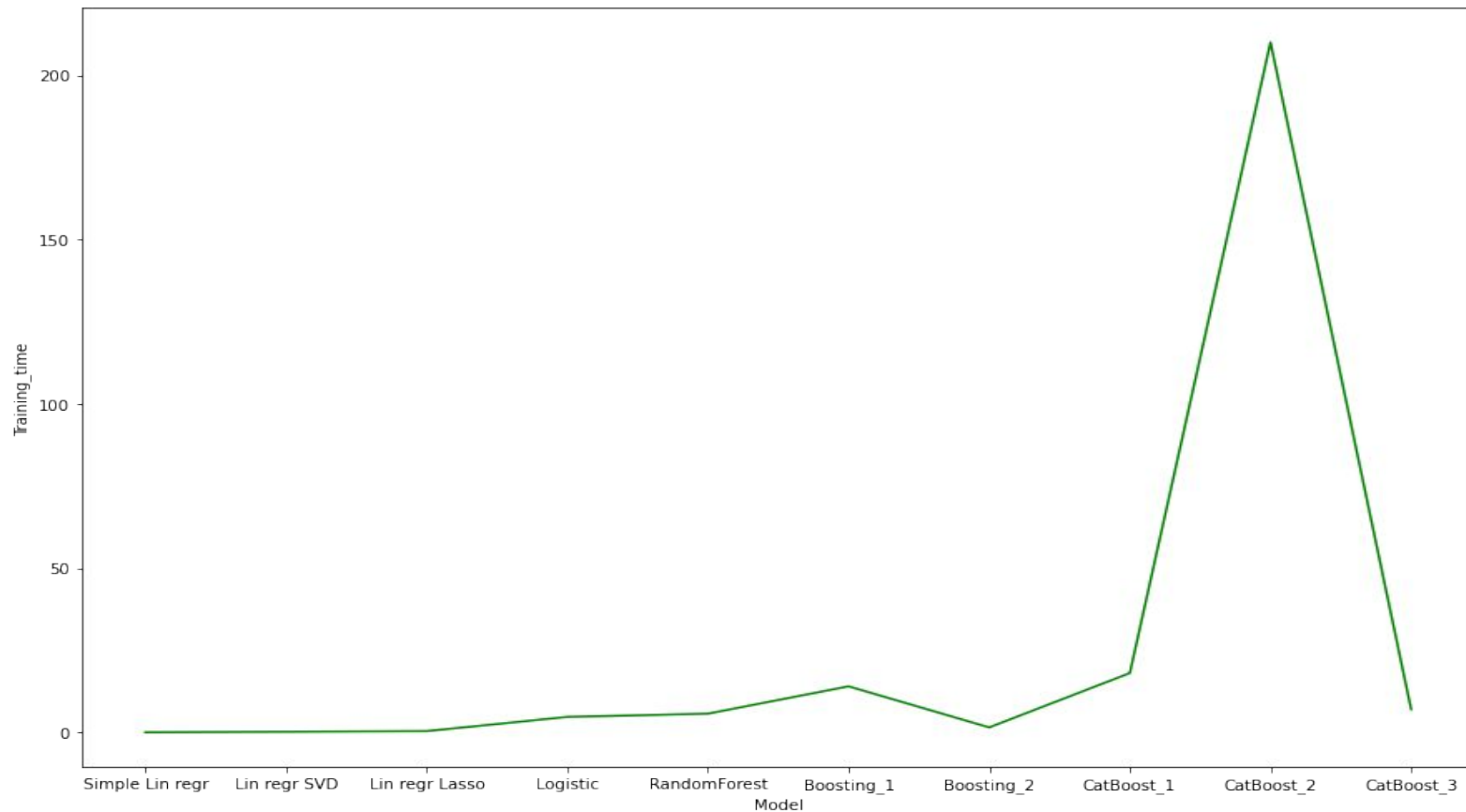
	Описание	Схема	Результат
CatBoost №3	Катбуст на стандартизованных признаках. Без доп обогащения признаками.		0.12496



Сравнение моделей по RMSE:



Сравнение моделей по времени обучения



A blurred city street scene at dusk. In the background, tall skyscrapers are visible against a warm, orange-hued sky. A large, multi-story brick building with many windows is prominent on the right. In the foreground, a yellow taxi is driving across a crosswalk, and several pedestrians are walking. The overall image has a soft, out-of-focus quality.

Спасибо за внимание!