

Прогнозирование задержек авиарейсов с использованием методов регрессионного анализа машинного обучения

Саяпин Никита Анатольевич

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

В статье рассматривается задача прогнозирования времени задержки прилёта авиарейсов с использованием регрессионных методов машинного обучения. В качестве исходных данных применяется открытый набор данных Flight Delay Dataset 2024, содержащий свыше 7 миллионов записей о рейсах внутри США. Из исходного массива сформирована рабочая выборка объёмом 500 000 строк. В ходе исследования проведён разведочный анализ данных, осуществлена предобработка и кодирование признаков.

Ключевые слова: машинное обучение, регрессионный анализ, задержки авиарейсов, прогнозирование, анализ данных.

Prediction of Flight Arrival Delays Using Regression Machine Learning Methods

Sayapin Nikita Anatolyevich

Sholom-Aleichem Priamursky State University

Student

Abstract

This article addresses the problem of predicting flight arrival delay duration using regression machine learning methods. The study is based on the Flight Delay Dataset 2024, an open dataset containing over 7 million records of domestic US flights. A working sample of 500,000 rows was drawn from the full dataset. Exploratory data analysis, feature preprocessing, and encoding were performed.

Keywords: machine learning, regression analysis, flight delays, prediction, data analysis.

1. Введение

1.1 Актуальность

Задержки авиарейсов являются одной из наиболее острых проблем гражданской авиации, ежегодно причиняя значительный экономический ущерб авиакомпаниям, аэропортам и пассажирам. По данным Бюро транспортной статистики США (BTS), в 2024 году свыше 20% внутренних рейсов прибыли с задержкой. Суммарные прямые убытки авиакомпаний от задержек оцениваются в миллиарды долларов ежегодно; при этом косвенные

потери — репутационные, логистические, связанные с компенсациями пассажирам — многократно увеличивают эту цифру.

В условиях цифровизации отрасли методы машинного обучения (МО) открывают возможности для точного прогнозирования продолжительности задержки на основе оперативно доступных данных. Применение регрессионных моделей позволяет перейти от статистических оценок к индивидуальным прогнозам для каждого рейса, что имеет практическую ценность для систем поддержки принятия решений в авиации.

1.2 Обзор исследований

В работе П.Г. Полторацкого рассмотрены перспективы машинного обучения в электронной торговле. Выделены данные, используемые для прогнозирования спроса, а также произведена классификация применяемых моделей обучения [1]. В статье М.М. Камилова и М.Х. Худайбердиева исследуется процесс моделирования спроса на товары с применением методов машинного обучения. Каждый метод оценивается по точности прогноза, способности адаптироваться к изменениям на рынке и времени обучения. Целью исследования является определение наиболее эффективного метода для прогнозирования спроса на товары [2]. К.Н. Алексеев предпринял попытку интегрировать модели интеллектуального анализа данных и CRM и предложить новую модель интеллектуального анализа данных для CRM. Цель исследования состоит в попытке интегрировать модели интеллектуального анализа данных и CRM и предложить новую модель интеллектуального анализа данных для CRM [3].

Основной задачей работы А.В. Щербининой и А.В. Алжеева стало сравнение прогностической способности классической модели машинного обучения - ARIMA, как наиболее распространенной и хорошо изученной baseline модели, и ML модели на основе последовательной нейронной сети - в данном случае LSTM [4]. В статье Л.А. Лейфер и Е.В. Черной представлены результаты экспериментального исследования эффективности применения алгоритмов машинного обучения для оценки недвижимости на примере оценки квартир на вторичном рынке Нижнего Новгорода. При проведении указанных экспериментов были применены различные методы машинного обучения и использованы нейронные сети [5].

1.3 Цель исследования

Целью данного исследования является построение и сравнительный анализ восьми регрессионных моделей машинного обучения для прогнозирования продолжительности задержки прилёта авиарейсов в минутах на основе реальных данных о рейсах США за 2024 год.

2 Разведочный анализ

2.1 Описание датасета и первичный анализ

В качестве исходных данных используется набор Flight Delay Dataset 2024 (автор: Hrishit Patil), опубликованный на платформе Kaggle. Датасет

сформирован на основе ежемесячных отчётов системы TranStats Бюро транспортной статистики США путём объединения 12 месячных CSV-файлов в единый очищенный массив. Полный файл содержит 7 079 081 строку и 35 столбцов, описывающих плановые и фактические параметры рейса: время вылета и прилёта, задержки, дальность маршрута, авиакомпанию, аэропорты вылета и назначения, причины задержки и признак отмены рейса.

Для работы в среде Google Colaboratory из полного массива методом случайной выборки (`random seed = 42`) извлечено 500 000 строк. Такой объём обеспечивает статистическую репрезентативность и допустимое время обучения моделей в рамках облачной среды (Рис.1).

```
Using Colab cache for faster access to the 'flight-data-2024' dataset.  
Датасет загружен в: /kaggle/input/flight-data-2024  
Найденные CSV файлы: ['flight_data_2024_sample.csv', 'flight_data_2024_data_dictionary.csv', 'flight_data_2024.csv']  
Используем: /kaggle/input/flight-data-2024/flight_data_2024.csv  
Загружено: 7,079,081 строк  
Рабочая выборка: 500,000 строк
```

Рисунок 1 – Загрузка датасета

2.2 Анализ пропущенных значений

На этапе разведочного анализа проведена проверка пропущенных значений по всем столбцам. Целевая переменная `arr_delay` (задержка прилёта в минутах) содержит пропуски, соответствующие отменённым или отклонённым рейсам. Все строки с отсутствующим значением целевой переменной были исключены из рабочей выборки. Пропуски в признаках `dep_delay` устранены аналогичным образом, поскольку данный признак включён в состав предикторов (Рис.2).

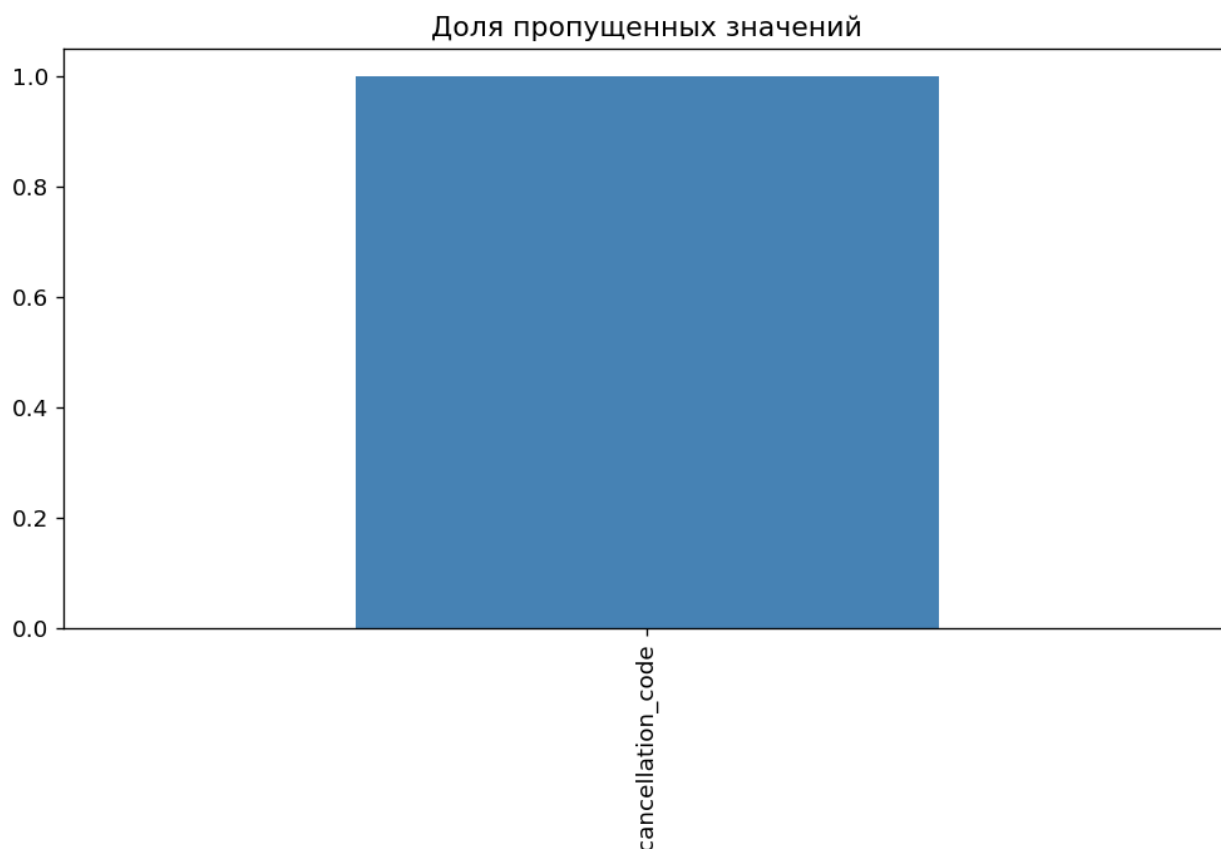
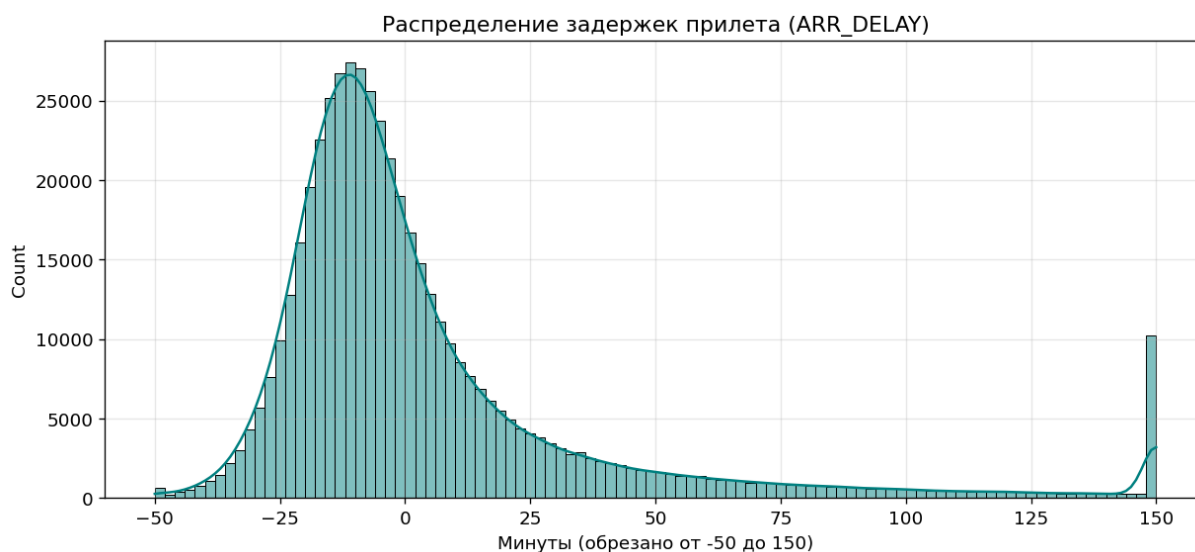


Рисунок 2 – Доля пропущенных значений по признакам

2.3 Распределение целевой переменной

Целевая переменная `arr_delay` представляет задержку прилёта в минутах (отрицательные значения соответствуют прилёту раньше расписания). Анализ распределения показал выраженную правостороннюю асимметрию: большинство рейсов прилетает близко к расписанию или с небольшим опережением, тогда как редкие случаи экстремальных задержек формируют длинный правый хвост распределения. Доля рейсов с задержкой прилёта свыше 15 минут составляет около 20% выборки (Рис.3).

Рисунок 3 – Гистограмма распределения задержки прилёта (`arr_delay`)

2.4 Анализ зависимостей

В рамках разведочного анализа изучены зависимости задержки прилёта от категориальных и числовых признаков. Анализ по дням недели показал, что наибольшая средняя задержка наблюдается в пятницу и воскресенье — в периоды пиковой загрузки воздушного пространства. Помесячный анализ выявил сезонную зависимость: наиболее высокие задержки характерны для июня–июля (летний туристический сезон) и декабря (рождественские праздники) (Рис.4).

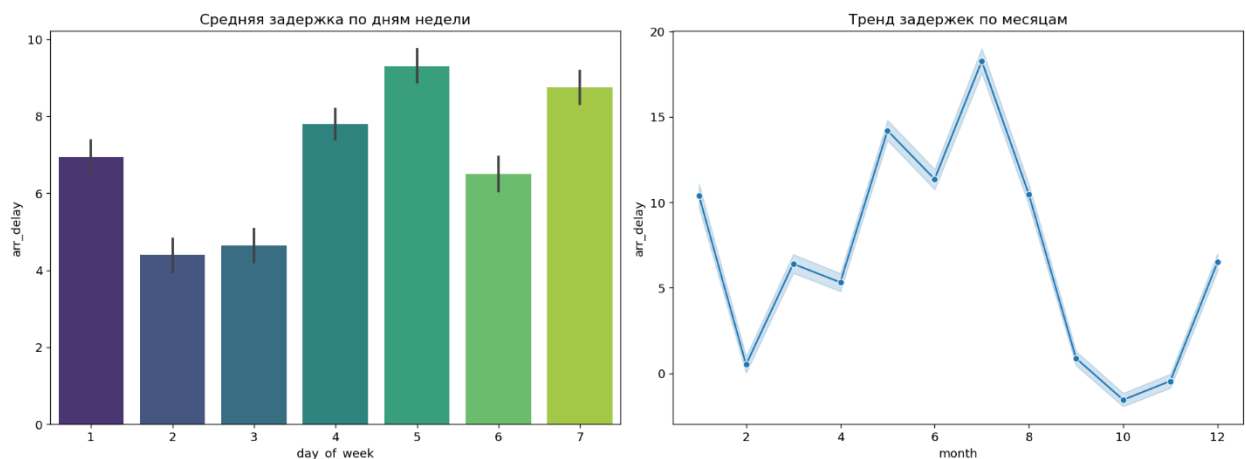


Рисунок 4 – Средняя задержка прилёта по дням недели и месяцам

Корреляционный анализ числовых признаков подтвердил сильную линейную связь между задержкой вылета (dep_delay) и задержкой прилёта (arr_delay), что физически объяснимо: рейс, вылетевший с опозданием, как правило, прибывает с соответствующей задержкой. Данная зависимость является ключевой для построения регрессионных моделей.

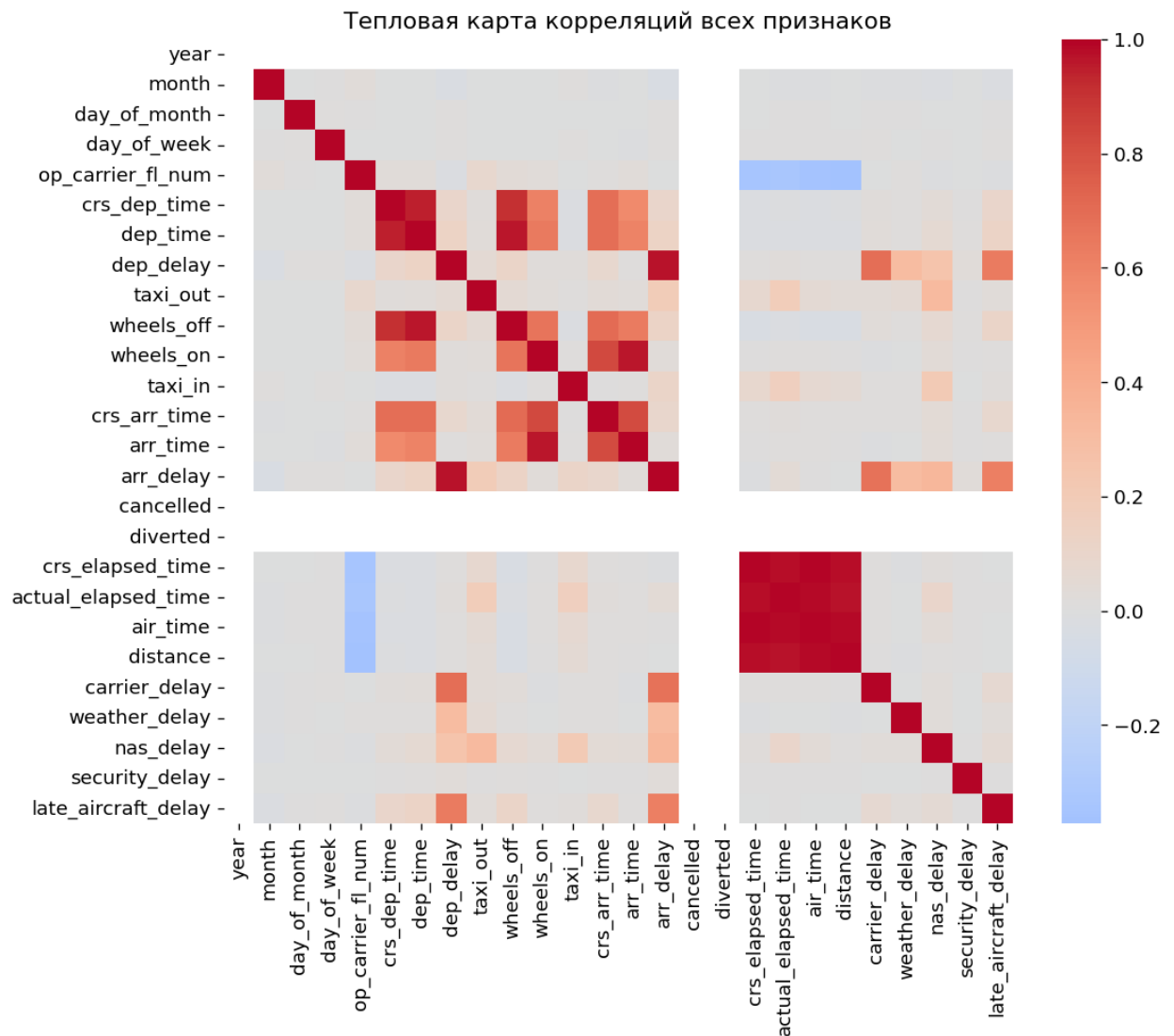


Рисунок 5 – Тепловая карта корреляций числовых признаков

Категориальные признаки `origin` и `dest` закодированы методом `LabelEncoder`. После удаления строк с пропусками в отобранных признаках итоговая рабочая выборка составила 491 973 строки и 5 признаков.

Данные разделены на обучающую и тестовую выборки в соотношении 80/20 (`random seed = 42`): обучающая — 393 578 строк, тестовая — 98 395 строк. Для линейных моделей и метода `k` ближайших соседей, чувствительных к масштабу признаков, применена стандартизация `StandardScaler`, обученная исключительно на обучающей выборке во избежание утечки данных (Рис.6).

```

from sklearn.preprocessing import LabelEncoder, StandardScaler

# Копируем данные для моделирования
df_prep = df_model.copy()

# Кодирование категориальных признаков
le = LabelEncoder()
for col in cat_features:
    df_prep[col] = le.fit_transform(df_prep[col].astype(str))

# Разделяем на признаки и целевую переменную
X = df_prep.drop(columns=[TARGET])
y = df_prep[TARGET]

# Разделяем на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=RANDOM_STATE)

# Масштабируем данные для линейных моделей и KNN
scaler = StandardScaler()
X_train_sc = scaler.fit_transform(X_train)
X_test_sc = scaler.transform(X_test)

print(f"Данные готовы: X_train {X_train.shape}, X_test {X_test.shape}")

Данные готовы: X_train (393578, 5), X_test (98395, 5)

```

Рисунок 6 – Разделение данных на обучающую и тестовую выборки

3. Построение и сравнение регрессионных моделей

В исследовании реализованы и сравнены восемь регрессионных алгоритмов машинного обучения из библиотеки `scikit-learn`.

Линейная регрессия (Linear Regression) — базовая модель, минимизирующая среднеквадратичную ошибку. Не имеет гиперпараметров регуляризации.

Ridge (гребневая регрессия) — линейная модель с L2-регуляризацией ($\alpha=1.0$), сглаживающей коэффициенты и снижающей влияние мультиколлинеарности.

Lasso — линейная модель с L1-регуляризацией ($\alpha=0.1$), обеспечивающей разреженность решения за счёт обнуления малозначимых коэффициентов.

Метод k ближайших соседей (KNN Regressor, $k=10$) — непараметрический метод, прогнозирующий значение как среднее по k наиболее близким объектам обучающей выборки.

Дерево решений (Decision Tree, $\text{max_depth}=10$) — нелинейная модель, рекурсивно разбивающая пространство признаков на прямоугольные области.

Случайный лес (Random Forest, 100 деревьев) — ансамбль деревьев решений, обученных на бутстрап-подвыборках с усреднением прогнозов.

Градиентный бустинг (Gradient Boosting, 100 итераций) — последовательный ансамбль, каждая итерация которого минимизирует остаточную ошибку предыдущего шага.

Экстремально рандомизированные деревья (Extra Trees, 100 деревьев) — ансамбль, отличающийся случайным выбором порогов разбиения, что обеспечивает более высокую скорость обучения по сравнению со случайным лесом (Рис.7).

```

model_instances = {
    'Linear Regression': LinearRegression(),
    'Ridge': Ridge(alpha=1.0),
    'Lasso': Lasso(alpha=0.1),
    'KNN Regressor': KNeighborsRegressor(n_neighbors=10, n_jobs=-1),
    'Decision Tree': DecisionTreeRegressor(max_depth=10, random_state=RANDOM_STATE),
    'Random Forest': RandomForestRegressor(n_estimators=100, n_jobs=-1, random_state=RANDOM_STATE),
    'Gradient Boosting': GradientBoostingRegressor(n_estimators=100, random_state=RANDOM_STATE),
    'Extra Trees': ExtraTreesRegressor(n_estimators=100, n_jobs=-1, random_state=RANDOM_STATE)
}

print(f"Инициализировано моделей для сравнения: {len(model_instances)}")

```

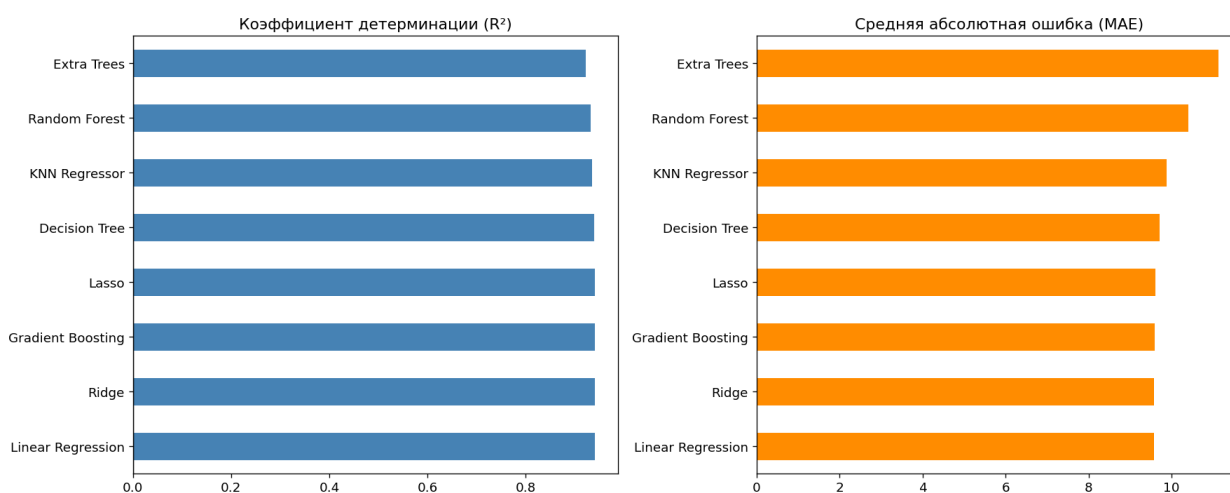
• Инициализировано моделей для сравнения: 8

Рисунок 7 – Инициализация регрессионных моделей

Для оценки качества моделей использованы три метрики (Рис.8): средняя абсолютная ошибка (MAE) — среднее отклонение прогноза от факта в минутах, интерпретируемое как типичная ошибка прогноза; коэффициент детерминации (R^2) — доля дисперсии целевой переменной, объяснённая моделью (1,0 — идеальный прогноз, 0 — модель равна среднему) (табл.1).

Таблица 1 – Сравнение регрессионных моделей по метрикам качества

№	Модель	MAE, мин	R^2	Ранг
1	Linear Regression	9,57	0,9412	1
2	Gradient Boosting	9,59	0,9410	2
3	Lasso	9,62	0,9408	3
4	Ridge	9,57	0,9412	1
5	Decision Tree	9,71	0,9393	5
6	KNN Regressor	9,87	0,9358	6
7	Random Forest	10,40	0,9317	7
8	Extra Trees	11,14	0,9216	8

Рисунок 8 – Сравнение моделей: графики R^2 и MAE

Все восемь моделей показали высокий коэффициент детерминации в диапазоне $R^2=0,9216-0,9412$. Наилучший результат по R^2 продемонстрировали линейная регрессия и Ridge ($R^2=0,9412$, MAE=9,57 мин), фактически разделившие первое место. Вплотную к лидерам примыкает градиентный бустинг ($R^2=0,9410$, MAE=9,59 мин).

Принципиально важным результатом является то, что простые линейные модели не уступили, а по ряду метрик превзошли сложные ансамблевые методы. Это свидетельствует о выраженном линейном характере связи между задержкой вылета и задержкой прилёта: самолёт, вылетевший с опозданием на N минут, в среднем прибывает также с задержкой, близкой к N минутам. Экипажи, как правило, не в состоянии полностью компенсировать задержку вылета в воздухе, что и порождает устойчивую линейную закономерность.

Наибольшую MAE среди всех моделей показал Extra Trees (11,14 мин), что связано с высокой степенью рандомизации при построении деревьев и, как следствие, большей дисперсией прогнозов. Случайный лес также показал MAE=10,40 мин при $R^2=0,9317$ — в данной задаче ансамблирование не принесло ощутимого прироста точности по сравнению с линейными методами.

Для лучшей модели (Linear Regression) построен график сравнения фактических и прогнозируемых значений задержки. Точки располагаются вблизи линии идеального прогноза ($y=x$), что визуально подтверждает высокое качество модели. Наибольшие отклонения наблюдаются в области экстремально больших задержек (свыше 200 минут), что объясняется редкостью таких событий в обучающей выборке (Рис.9).

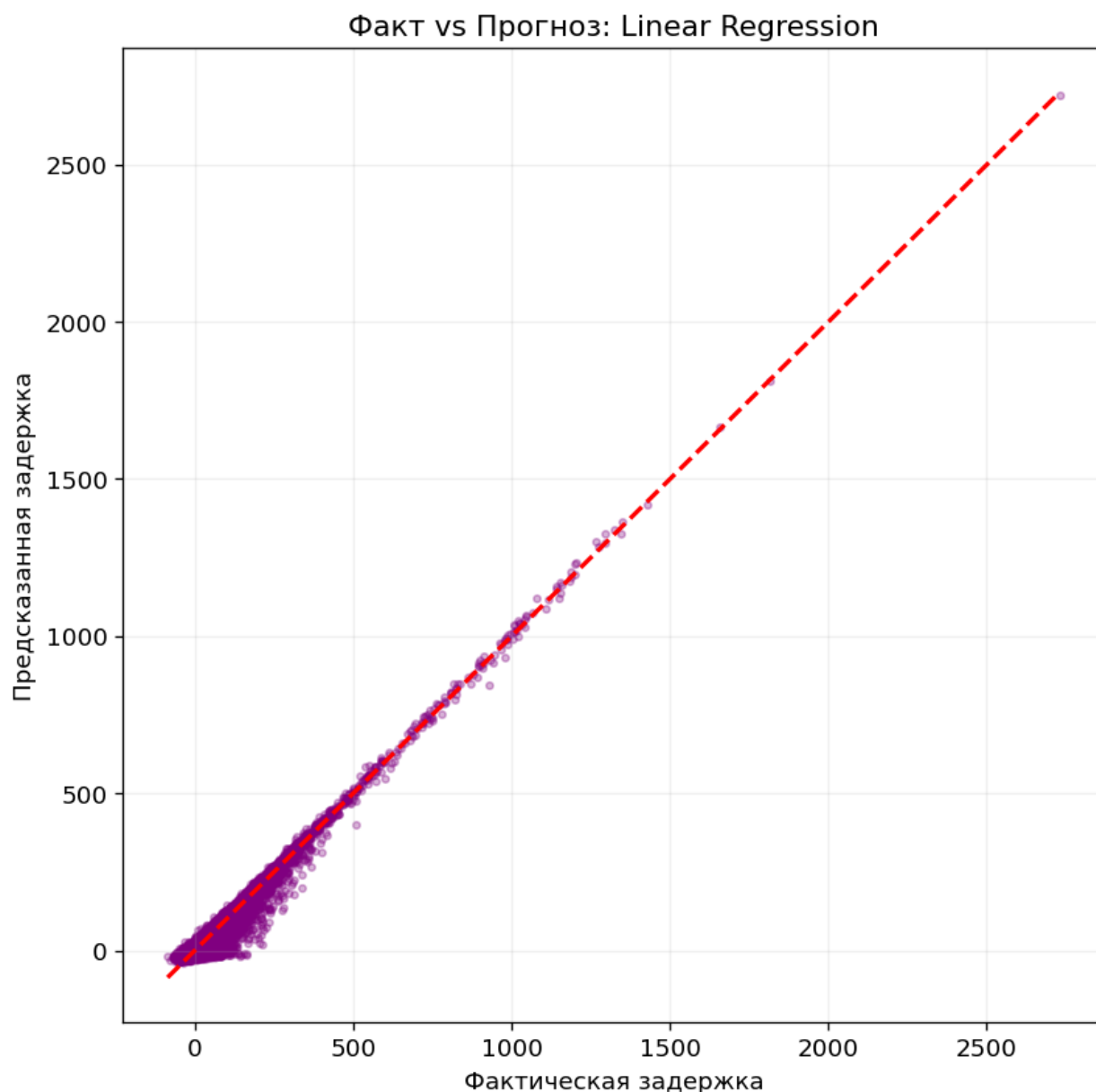


Рисунок 9 – Факт vs Прогноз: Linear Regression

4 Выводы

В данной работе проведён сравнительный анализ восьми регрессионных моделей машинного обучения для прогнозирования задержки прилёта авиарейсов. На основе реального набора данных объёмом 7 079 081 строк (рабочая выборка — 500 000 строк) из пяти признаков построены и оценены модели линейной регрессии, Ridge, Lasso, KNN, дерева решений, случайного леса, градиентного бустинга и Extra Trees.

Все модели продемонстрировали высокое качество прогнозирования ($R^2 > 0,92$). Лучший результат показали линейная регрессия и Ridge ($R^2 = 0,9412$, MAE=9,57 мин), вплотную к ним примыкает градиентный бустинг ($R^2 = 0,9410$). Тот факт, что простые линейные модели оказались наравне с ансамблевыми, является самостоятельным научным результатом: он подтверждает линейный характер зависимости задержки прилёта от задержки

вылета и ставит под сомнение целесообразность применения вычислительно затратных ансамблевых методов в данной постановке задачи.

Полученные результаты могут быть использованы при разработке систем информирования пассажиров, оптимизации расписания рейсов и систем поддержки принятия решений авиакомпаний. Направлениями дальнейшего исследования являются включение погодных данных в состав признаков, применение методов временного ряда (LSTM, Prophet) и построение классификационной модели для задачи бинарного прогнозирования задержки.

Библиографический список

1. Полторацкий П. Г. Прогнозирование спроса в электронной коммерции методами машинного обучения // Вестник науки. 2024. №6 (75).
2. Камилов М. М., Худайбердиев М. Х., Алимжанова О. С. К. Процесс моделирования спроса на товары с использованием алгоритмов машинного обучения // *Al-Farg'oniy avlodlari*. 2024. №. 2. С. 249-254.
3. Алексеев К. Н. Организация CRM-аналитики с использованием технологии Data mining // Финансовые рынки и банки. 2021. №. 4. С. 4-9.
4. Щербина А. В., Алжеев А. В. Сравнительный анализ качества прогнозирования классической статистической модели и модели машинного обучения на данных российского фондового рынка // Ученые записки Российской академии предпринимательства. 2021. Т. 20. №. 3. С. 52-63.
5. Лейфер Л. А., Чёрная Е. В. Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир // Имущественные отношения в Российской Федерации. 2020. №. 3 (222). С. 32-42.