УДК 004.8

ПОСТРОЕНИЕ МОДЕЛИ ПРЕДСКАЗАНИЯ ДЛИТЕЛЬНОСТИ ЗАДЕРЖЕК АВИАРЕЙСОВ

С. С. Ноздрин, И. Л. Каширина

Воронежский государственный университет

Аннотация. Статья представляет собой исследование использования современных методов анализа данных и машинного обучения для предсказания времени задержек авиарейсов на основе исторических данных. В результате выполненной работы была построена модель, которая решает задачу прогнозирования с заданной степенью точности.

Построенная модель может быть интересна руководителям авиакомпаний и аэропортов, а также исследователям, изучающим современные методы предиктивной аналитики.
Ключевые слова: анализ, машинное обучение, авиарейсы, модель.

Введение

В современном мире многие компании нередко прибегают к технологиям анализа данных и машинного обучения для оптимизации работы их сервисов и построению прогнозов, исключением не являются и авиакомпании.

Получение своевременной информации о возможных задержках рейсов дает возможность компании перестроить работу в аэропорту, избежать внештатных ситуаций, а также заблаговременно предоставить необходимую информацию пассажирам.

Цель данной работы: взяв за основу исторические данные (набор для обучения) по пунктуальности авиарейсов за период с октября 2015 года по сентябрь 2018 года, необходимо разработать модель машинного обучения, которая будет прогнозировать длительность задержки рейсов по отправлению (в минутах) на представленных тестовых данных. Рейс считается задержавшимся, если время его фактического отправления больше времени отправления по расписанию. Если рейс вылетает раньше расписания, задержка считается равной нулю.

К разрабатываемому проекту выдвинуты следующие требования:

1. Получить наименьшее значение метрики RMSE на тестовом наборе данных. Имеющиеся в наличии решения обеспечивают значение метрики 26.9 на проверочных данных, не участвующих в обучении. По результатам проведенного анализа необходимо улучшить имеющийся результат.

2. Аналитическую модель необходимо разрабатывать с использованием Python версии 3.5 и выше.

В данной работе для решения задачи нами были применены такие методы, как случайный лес, xgboost, метод кластеризации k-средних [1]. Оценка и сравнение построенных классификационных алгоритмов производились с помощью метрики RMSE.

В настоящее время исследованию пунктуальности рейсов и разработки предсказательной модели большее значение уделяется зарубежом, поэтому стоит обратиться к иностранным исследователям. Большая часть работ по изучению данной задачи сводилась к исследованию причин задержек рейсов и построению модели классификации с целью определения наличия задержки как таковой.

Так, команда исследователей из Португалии во главе с Нуно Фернандешом в работе «Factors influencing charter flight departure delay» [2], изд. с 2020 г. получила важные признаки, необходимые для предсказания наличия задержки: информация об отмене предыдущего полета данного рейса, продолжительность рейса, количество двигателей самолета, ширина и долгота расположения аэропортов прибытия и отправления. Попытки предсказать длительность за-

держек у данных исследователей не принесли успехов и ограничились построением модели для предсказания рейсов с задержкой не более, чем 60 минут.

Также египетскими учеными в работе «Machine learning techniques for analysis of egyptian flight delay» [3], изд. с 2018 г. были предприняты аналогичные попытки по построению модели классификации задержек рейсов местных авиалинии. Значимыми признаками для данного классификатора оказались продолжительность полета, тип самолета, время вылета и прилета.

Таких образом, многие иностранные исследователи занимаются анализом задержек рейсов и строят предиктивные модели, получая наиболее важные факторы для предсказания. Однако, большинство этих моделей сводятся к классификации рейсов, а не построению регрессионой модели, т. е. предсказанию длительности задержки.

Материалы и методы

Исходная выборка

Задача, обсуждаемая в данном исследовании, поставлена компанией Рамакс, которая предоставила неперсонифицированные данные о рейсах с 2015 по 2018 год. В исходную выборку были отобраны наиболее значимые по мнению компании признаки для предсказания, которые представлены в табл. 1.

Тренировочная выборка вмещает в себя 675259 образцов, что соответствует количеству рейсов, а тестовые данные содержат информацию о 65429 рейсах.

Таблица 1

Исходные признаки

Категория признаков	Переменные
Непрерывные	Дата Рейса, Время отправления по расписанию, Номер ВС, Пассажиры факт Всего, Время прибытия по расписанию, Пассажиры факт J, Пассажиры факт W ,Пассажиры факт Y.
Категориальные	Рейс, Тип рейса, А/П отправл, А/П прибыт, Стоянка отправл, Терминал приписки (отправл), Тип ВС Гейт прибыт, Статус.

Особенностью данных являлось распределение зависимой переменной. На рис. 1 можно заметить, что в основном задержек не случалось или они совсем маленькие и находятся в окрестности нуля. 391069 из 675259 образцов имели 0 задержку, более подробная информация представлена в таблице ниже.

Также отметим, что 0.95 квантиль — это задержки от 40 минут и меньше, а 0.99 квантиль — 160 минут и меньше. Поэтому для наглядности на рис.1 гистограмма построена с учетом 0.95 квантиля, для решения же задачи большие задержки не удалялись. Сложность данной задачи как раз проявлялась в трудности обнаружения больших задержек.

Таблица 2

Статистика зависимой переменной

Характеристики	Время задержки по отправл, мин.
Количество	675259
Среднее	9.75
Стан.отклонение	43.44
Мин.	0
1 квартиль	0
2 квартиль	0
3 квартиль	5
Макс	1436

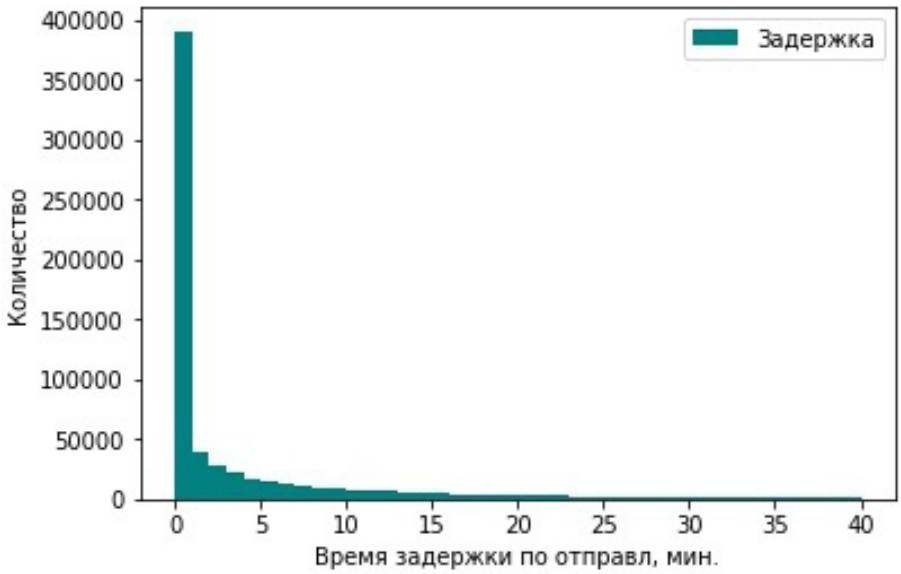

Рис. 1. Гистограмма распределения зависимой переменной

На основе исходных данных были построены такие дополнительные признаки, как величина средних задержек по всем категориальным переменным (средняя задержка по каждому типу рейса, номеру вс, и.т. д). Важным признаком оказался месяц полета. На рис. 2 видно, что зимой средняя задержка возрастает в 1,5–2 раза относительно других сезонов.

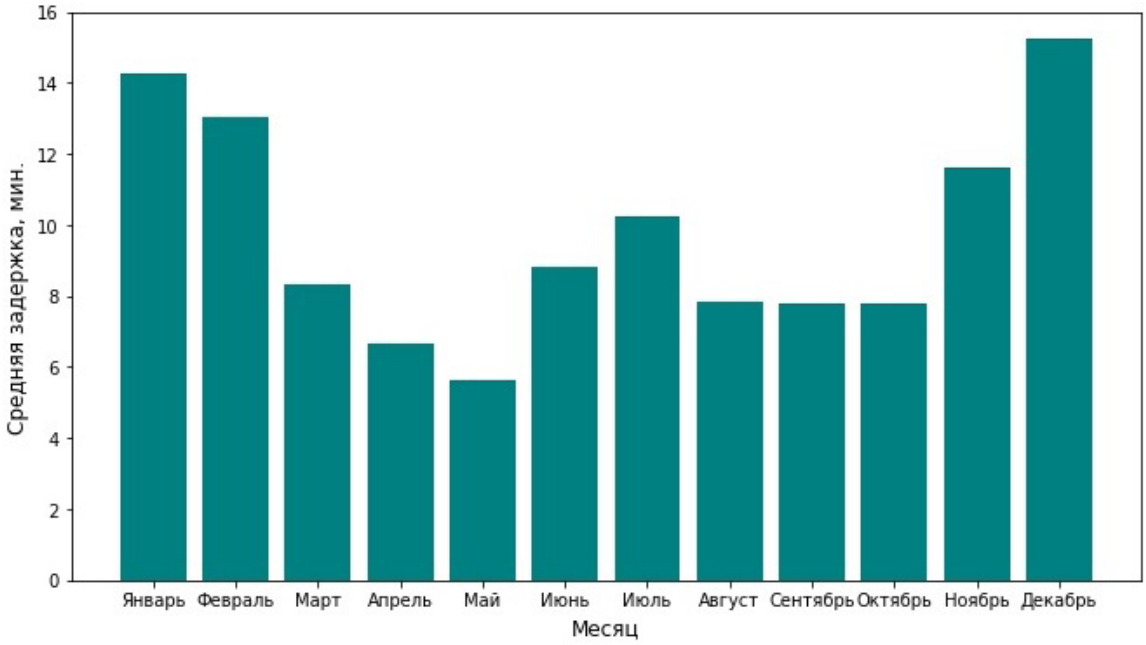

Рис. 2. Средняя задержка по месяцам

Также анализировались и сезонные факторы, например, влияние дней недели, времени суток, даты на длительность задержки. Матрица корреляций между новыми количественными переменными и зависимой переменой представлена на рис. 3.

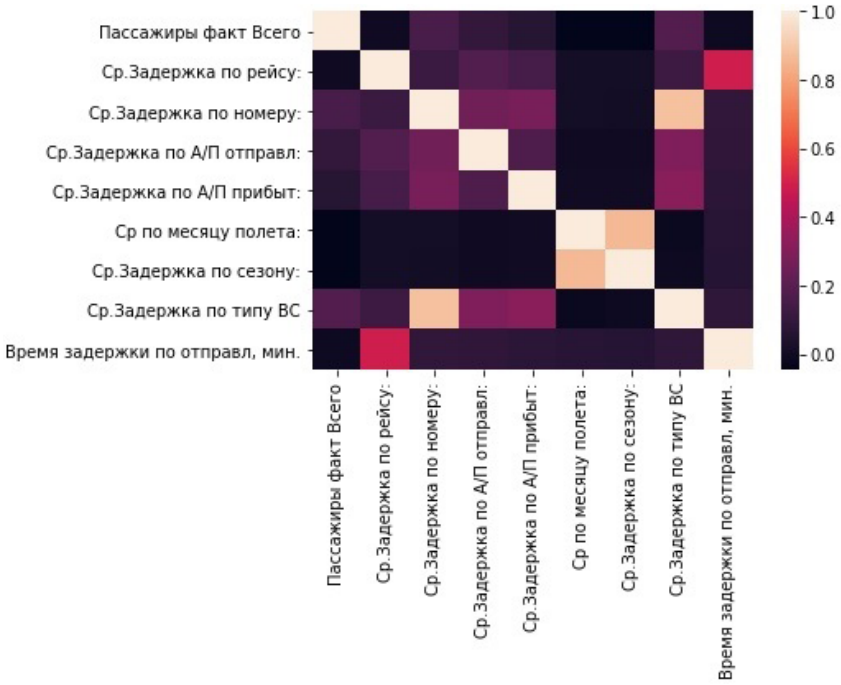

Рис. 3. Матрица корреляций новых признаков с зависимой переменной

Построение моделей

Первичный анализ включал кластрезацию исходных данных. Для кластеризации использовался алгоритм k-средних. Это очень популярный и простой метод, позволяющий разделить исходные данные на несколько групп.

Наилучшим разбиением считается то, которое минимизирует величину V из формулы (1), где S_i — это i-й кластер, x — элемент из данного кластера, а m_i — центр кластера.

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - m_i)^2. \quad (1)$$

Оптимальное количество кластеров определялось с помощью графика каменистой осыпи, представленного на рис. 4. На основе этого графика первоначально было принято решение об использовании 6 кластеров.

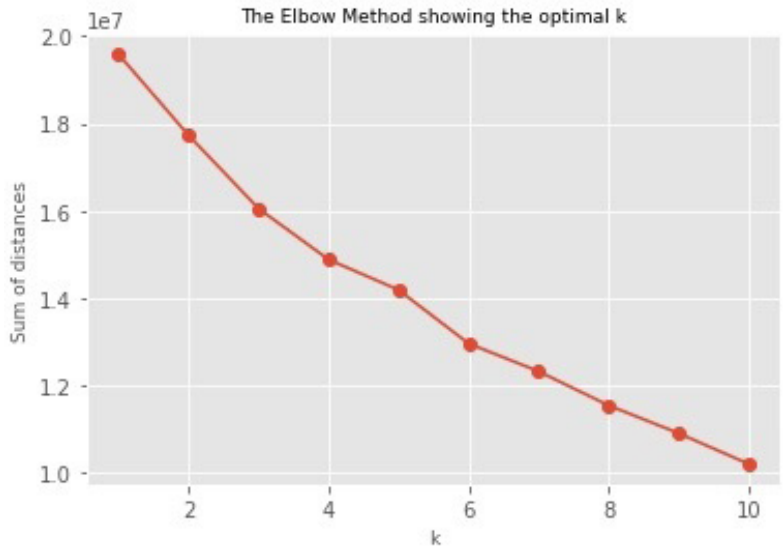

Рис. 4. График каменистой осыпи

Впоследствии некоторые кластеры были объединены в один, т. к. средняя задержка в этих кластерах была схожей. Распределение средних задержек по кластерам представлено в таблице ниже.

Таблица 3

Средние задержки по кластерам

Номер кластера	Средняя задержка	Размер кластера
1	7.315	170695
2	15.033	88166
3	12.216	100688
4	12.831	54511
5	6.065	260246
6	517.086	973

В итоге, модели машинного обучения строились для 4 кластеров: маленьких задержек(кластеры 1 и 5), небольших(кластры 3 и 4), средного(2) и большого(6). Важным итогом кластерного анализа являлся результат получения явного кластера с большими задержками, что относится к безусловным плюсам данной модели.

Для большей части кластеров использовался метод машинного обучения - xgboost (модифицированная версия градиентного бустинга).

В основе XGBoost[6] лежит алгоритм градиентного бустинга деревьев решений. Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Обучение ансамбля проводится последовательно. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке. Следующая модель, которая будет добавлена в ансамбль, будет предсказывать эти отклонения. Таким образом, добавив предсказания нового дерева к предсказаниям обученного ансамбля, мы можем уменьшить среднее отклонение модели, которое является целью оптимизационной задачи. Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил «ранней остановки». Данная модель имеет несколько гиперпараметров, которые необходимо подбирать для каждой конкретной задачи.

Нахождение оптимальных ответов последующего дерева исходит из решения оптимизационной задачи нахождения минимума функции (3) по параметру w

$$L = l(y_i, \hat{y}_i^{(t-1)} + w_t(x_i)) + \gamma T + 0.5\lambda \sum_{j=1}^{T} w_j^{\,2}, \tag{2}$$

где l — функция потерь, $y_i, y_i^{(t-1)}$ — значение i-го элемента обучающей выборки и сумма предсказаний первых t деревьев соответственно, x_i — набор признаков i-го элемента обучающей выборки, $w_t(x_i)$ — предсказание на i-м элементе обучающей выборки, γT — регуляризационное слагаемое, штрафующее за чрезмерное количество узлов в дереве (T), а λ — коэффициент L_2 регулязации.

Также, для одного из кластеров наилучшим образом показала себя модель случайного леса.

Случайный лес — алгоритм машинного обчения, заключающийся в использовании ансамбля решающих деревьев. Алгоритм применяется как для задач классификации, так и для задач регрессии или кластеризации. Каждое решающее дерево само по себе дает не очень высокое качество классификации, но за счет их большого количества результат выходит хорошим. Наиболее распространенный способ построения деревьев — бэггинг. Идея заключается в генерации случайной подвыборки с повторениями. Далее строится решающее дерево, которое классифицирует образцы данной подвыборки, причем в ходе создания очередного узла дерева

выбирается набор признаков, на основе которых производится разбиение, из которых выбирается наилучший (например, с использованием критерия Джини). Дерево строится до полного исчерпания подвыборки. В задачах регрессии предсказывается значение путем усреднения ответов деревьев.

Также одной из важных стадий построения моделей машинного обучения являлось формирование новых признаков и удаление менее информативных из исходной выборки. Для каждой из групп модель строилась по-своему, индивидуальному признаковому пространству. В табл. 4 представлены признаки, которые оказались наиболее полезными для решения данной задачи.

Таблица 4

Используемые признаки

Категория признаков	Переменные
Из исходной выборки	Пассажиры факт Всего, Тип рейса, Тип ВС Стоянка отправл, Год полета, День полета
Новые	А/П прибыт – SVO, Рейс с 0 задержкой Средняя задержка по рейсу, Праздничный день Год экспл, Сезон полета

На основе построенных моделей был сделан вывод о том, какие признаки оказались наиболее важными, на рис. 5 приведена столбчатая диаграмма. Другие типы столбцов, имеющие значение важности менее 0.05, были отброшены, т. к. существенно не улучшали качество модели.

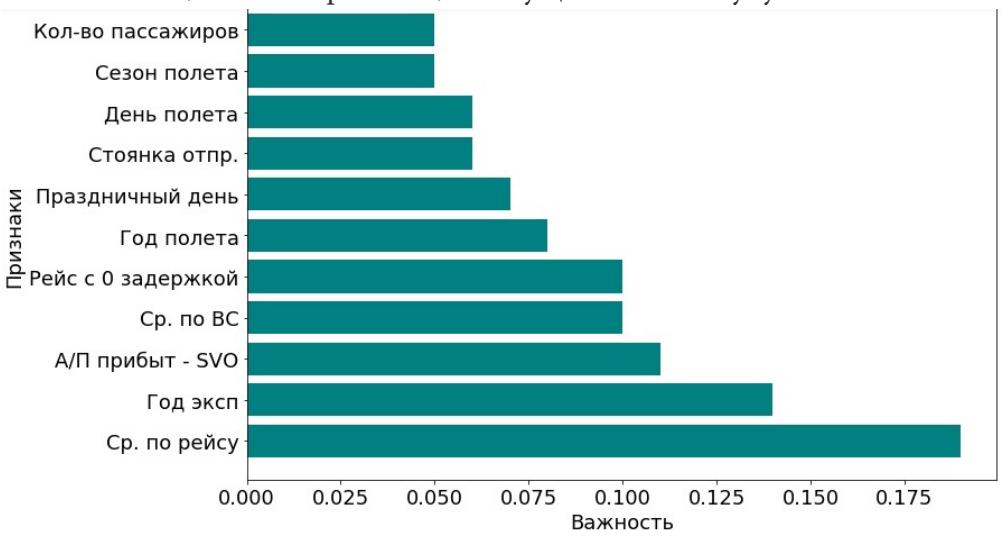

Рис. 5. График важности признаков

Помимо исходных признаков к общему пространству были добавлены новые признаки, построенные на основе предыдущих. Данные признаки оказались полезными, так, например, в праздничные дни и их предверие повышается спрос на рейсы, а также в данных существовала особая группа рейсов, которая практически никогда не задерживается.

Не менее значимым оказался признак сезон полета, т. к. зимой средняя задержка возрастает в несколько раз, что заметно по рис. 2. Также на основе данных было отмечено, что рейсы, прибывающие в аэропорт Шереметьево(SVO) чаще задерживаются на продолжительное время.

Также дополнительно собиралась и подключалась к исходным данным информация, связанная с годом введения в эксплуатацию типа ВС. Это важный признак, т. к. чем старше воздушное средство, тем тщательнее его проверяют перед взлетом, чаще возникают внештатные результаты перед полетом.

Результаты и их обсуждения

В качестве показателя производительности модели была использовала метрика RMSE, которая является стандартной для задач регрессии.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}.$$

Данный показатель считается по выборке длины n, y_i — реальное значение предсказываемой переменной, а $\hat{y}_i$ — значение, предсказанное моделью.

К достоинствам данного показателя относят легкость интерпритации, а также дифференцируемость и как следствие, возможность использования данной метрики как функции ошибок для методов, основанных на нахождении минимума таких показателей. Главным недостатком же является высокая чувствительность RMSE к выбросам, т. е. большим отклонениям разности в числителе. Даже редкие такие случаи могут привести к серьезному увеличению данной метрики.

Таблица 5

Результаты построенных моделей

Номер кластера	Модель	Показатель RMSE
1	xgboost	22.30
2	xgboost	33.98
3	Random Forest	25.62
4	xgboost	310.92

Выбор наилучшей модели для каждой из групп(кластеров) осуществлялся с помощью показателя метрики RMSE, а наилучшие гиперпараметры подбирались с помощью техники кросс-валидации. В итоге получилось улучшить значение данной метрики до 26.45, что является хоть и не таким существенным, как хотелось бы, но улучшением исходного показателя (26.9).

В ходе работы для решения задачи регрессии — нахождения длительности задержки авиарейсов использовался кластерный анализ, позволивший разделить данные на несколько групп, для каждой из которых строились модели машинного обучения, такие как случайный лес и xgboost, со своим признаковым пространством.

Важной особенностью задачи является то, что больших задержек очень мало, их очень тяжело идентифицировать и точно предсказать, а метрика RMSE очень чувствительна к большим отклонениям.

Таблица 6

Результаты построенных моделей

	С учетом посл.кластера	Без учета посл.кластера
Показатель RMSE	26.45	25.66

Модель справляется с тем, чтобы предсказывать маленькие задержки и находить большие, но точность прогнозов в кластере с крупными задержками оставляет желать лучшего. Именно это и способствует резкому увелечению метрики. Сравнение данного показателя без учета посл. кластера (включающего всего 27 образцов из общего количества в 65429 образцов в тестовом множестве) и результата модели приведены в таблице выше.

Заключение

В результате данного исследования был проведен всесторонний анализ задержек рейсов и их предикторов. Для решения поставленной задачи были применены следующие методы: алгоритм кластеризации K-средних, случайный лес и xgboost. В ходе работы был построен ряд моделей, позволивший незначительно улучшить исходную метрику данной регрессионой задачи. Результат был достигнут путем разбиения исходных данных на кластеры и применения техники кросс-валидации [4], позволившей найти лучшие гиперпараметры для моделей, давшие наилучшие показатели метрики RMSE.

Литература

1. Scikit-learn.cluster.KMeans – машинное обучение на Python. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans (дата обращения: 10.12.2019).

2. *Fernandes, N.* Factors influencing charter flight departure delay / N. Fernandes, S. Moro, C. J. Costa, M.Apracio // Academia. – 2018. – № 1. – URL: https://www.academia.edu/44018330/Factors-influencing-charter-flight-departure-delay (дата обращения: 10.10.2020).

3. *Al-Tabbakh, S. M.* Machine learning techniques for analysis of egyptian flight delay / S. M. Al-Tabbakh, H. M. Mohamed, H. El-Zahed // Academia. – 2020. – № 1. – URL : https://www.academia.edu/36835689/MACHINE-LEARNING-TECHNIQUES-FOR-ANALYSIS-OF-EGYPTIAN-FLIGHT-DELAY (дата обращения: 10.10.2020).

4. *Шолле, Ф.* Глубокое обучение на Python / Ф. Шолле. – СПб : Питер, 2018. – 400 с.

5. Scikit-learn – машинное обучение на Python. – URL: https://github.com/scikit-learn (дата обращения: 15.01.2020).

6. xgboost – библиотека машинного обучения. – URL: https://xgboost.ai (дата обращения: 27.01.2020).