

УНИВЕРСИТЕТ НАУКИ И ТЕХНОЛОГИЙ «МИСИС»
ИНСТИТУТ КОМПЬЮТЕРНЫХ НАУК
НАПРАВЛЕНИЕ 09.03.00

Курсовая работа

По дисциплине «Прикладной статистический анализ» на тему:

Разработка модели прогнозирования инцидентов в авиации

Работу выполнил:

Студент 3 курса

Группы БИВТ-21-5

Савосин А. А.

Научный руководитель:

Маркарян А. О.

Москва, 2023 г.

Содержание

Содержание.....	2
Введение	3
1. Анализ характеристик объекта исследования	4
1.1 Описание объекта исследования	4
1.2 Анализ объекта исследования с помощью статистических показателей	4
1.3 Выявление причинно-следственных связей	5
1.4 Постановка задачи моделирования	6
2. Моделирование статистических зависимостей	7
2.1 Формализация и классификация переменных	7
2.2 Проверка гипотезы о нормальном распределении выходной величины	7
2.3 Корреляционный анализ.....	8
2.4 Построение регрессионной модели.....	9
2.4.1 Структурная идентификация модели	9
2.4.2 Параметрическая идентификация модели.....	10
3. Исследование модели	11
3.1 Анализ статистической значимости уравнения регрессии.....	11
3.2 Анализ статистической значимости коэффициентов уравнения регрессии.....	12
3.3 Исследование мультиколлинеарности факторов.....	13
3.4 Применение шагового регрессионного анализа для улучшения модели	13
4. Программная реализация и численное исследование результатов моделирования .	15
4.1 Обоснование выбора и описание программного обеспечения	15
4.2 Описание основных модулей программы.....	15
4.3 Численное исследование результатов моделирования.....	19
4.4 Улучшение качества предсказания с помощью методов машинного обучения .	19
Выводы	21
Приложения	22

Введение

Авиационная безопасность является приоритетным вопросом в современном мире, поскольку инциденты в авиации могут иметь серьезные последствия, включая потерю человеческих жизней. Оценка и прогнозирование возможных летальных исходов при авиационных происшествиях играют критическую роль в предоставлении эффективной медицинской помощи и организации спасательных операций на месте происшествия.

С увеличением объемов воздушного транспорта и ростом пассажиропотока авиационные инциденты становятся все более сложными и требуют детального анализа для разработки эффективных стратегий предотвращения и реагирования. Актуальность данного исследования обусловлена необходимостью предсказания возможных летальных исходов с целью максимально оперативного и эффективного привлечения медицинских и спасательных ресурсов.

Целью данной работы является разработка модели прогнозирования количества летальных исходов при авиационных инцидентах. Достижение данной цели предполагает решение следующих задач: анализ характеристик объекта исследования, моделирование статистических зависимостей, исследование модели и программная реализация.

Объектом исследования выступают авиационные инциденты, а предметом – количественные и качественные показатели летальных исходов, связанных с данными инцидентами. Исследование направлено на выявление закономерностей, позволяющих эффективно прогнозировать потенциальные последствия и принимать оперативные меры по предоставлению медицинской помощи и спасательных операций.

1. Анализ характеристик объекта исследования

1.1 Описание объекта исследования

Собранная информация с интернет-архива авиационных инцидентов [1], представляет собой набор данных о катастрофах в авиации с 1918 года по 2022 год. Подобный набор содержит всю полезную информацию о каждом авиационном крушении: время происшествия, локацию, характеристики летального аппарата, количество погибших и т.д. В данном исследовании такой набор можно называть генеральной совокупностью, ведь первое крушение самолёта произошло 17 сентября 1908 года. Объект исследования – общее количество летальных исходов во время инцидента, является ничем иным, как набором целых чисел длиной 28536 значений.

1.2 Анализ объекта исследования с помощью статистических показателей

Для набора значений летальных исходов были вычислены основные статистические показатели.

Средняя – среднее арифметическое наблюдаемых значений, вычисленное по формуле $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$, где N – количество значений. Среднее арифметическое в нашем случае равно 5.5674, это значит, что в среднем фиксируются 6 случае летального исхода при авиационном происшествии.

Дисперсия – мера рассеивания значений относительно средней, вычисляется по формуле $D(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$, где N – количество значений. Дисперсия примерно равна 279.3312.

Среднее квадратическое отклонение, вычисленное по формуле $\sigma(X) = \sqrt{D(X)}$ равняется 16.71. Именно на столько может отклониться случайное значение из совокупности относительно средней величины. Полученное значение довольно велико.

Коэффициент вариации – выраженное в процентах отношение среднего квадратического отклонения к средней: $V(X) = \frac{\sigma(X)}{\bar{x}}$. В данном случае коэффициент равен 300%, что свидетельствует о высокой вариации данных. Это значит, что применение средней для обобщения показателей совокупности нецелесообразно.

Размах – разница минимального и максимального значений совокупности. Равно 520, при минимальном значении 0 и максимальном 520.

Модой называется значение с наибольшей частотой. В данном случае мода равняется 0 с частотой 11851. Это значит, что 41.5% всех инцидентов прошли без потерь.

Медиана – значение, которое делит отсортированный ряд на две равные части. Если число вариант нечетное, то $Me = x_{m+1}$. Если число вариант четное, то $Me = \frac{x_m + x_{m+1}}{2}$, где m – индекс центрального элемента. В текущем случае медиана равна 1, это также подтверждает теорию о том, что большинство аварий проходит с крайне малым количеством потерь.

1.3 Выявление причинно-следственных связей

Исследование причинно-следственных связей в авиационных инцидентах и количестве смертей представляет собой сложную задачу, требующую комплексного анализа нескольких важных факторов. Один из ключевых аспектов – техническое состояние воздушного судна. Старение самолета и его общее техническое состояние имеют непосредственное влияние на вероятность возникновения аварии. Более того, тип воздушного судна также оказывает важное воздействие, учитывая различия в стандартах безопасности для пассажирских, грузовых и военных полетов.

Фаза полета является критическим моментом, определяющим характер возможных аварий и степень их тяжести. Например, аварии при взлете или посадке могут иметь более серьезные последствия для экипажа и пассажиров, чем инциденты, происходящие на крейсерском полете. Также необходимо учитывать тип полета, поскольку различия между пассажирскими рейсами, грузовыми перевозками и военными миссиями вносят свои особенности в степень риска и последующие последствия.

Место крушения играет важную роль в контексте выживаемости. Ландшафт, климатические условия и близость к населенным пунктам могут влиять на количество жертв и успешность операций по спасению. Год выпуска воздушного судна также является значимым фактором, поскольку старые самолеты более подвержены техническим сбоям, что увеличивает риск инцидентов и, как следствие, количество смертей.

Анализ безопасности воздушного пространства и стандартов в авиационной промышленности отдельных стран и регионов также является неотъемлемой частью исследования. Различия в подходах к безопасности могут быть связаны с уровнем инвестиций в обучение экипажа, техническое обслуживание, а также общую культуру безопасности.

Особое внимание уделяется причинам крушения. Технические сбои, человеческий фактор, погодные условия и террористические акты представляют собой основные источники риска. Анализ этих причин помогает выявить корреляции между ними и последующим количеством смертей.

В итоге, комплексный анализ вышеперечисленных факторов не только способствует лучшему пониманию причинно-следственных связей в авиационных инцидентах, но и позволяет выделить ключевые области для улучшения безопасности в авиации.

1.4 Постановка задачи моделирования

Постановка задачи моделирования направлена на разработку и обучение модели, способной предсказывать количество смертей в результате авиационных инцидентов. Для достижения этой цели предполагается использование специального набора данных, содержащего информацию о различных параметрах авиационных происшествий, таких как техническое состояние воздушных судов, тип полета, фаза полета, место крушения, год выпуска и другие ключевые переменные.

Первоочередной задачей является подготовка и очистка данных, а также определение признаков, имеющих наибольшее влияние на количество смертей. Для эффективного моделирования необходимо также провести анализ структуры данных, выявить возможные пропуски или выбросы, которые могут повлиять на качество модели.

Следующим этапом является выбор подходящего алгоритма, способного учесть особенности предсказания количества смертей в зависимости от различных параметров. Обучение модели будет проводиться на обучающем наборе данных, а затем ее эффективность будет проверена на тестовой выборке.

Оценка качества модели включает в себя анализ ее точности, чувствительности и специфичности, а также других метрик, адаптированных к конкретной задаче предсказания количества смертей при авиационных инцидентах.

Основной целью данного моделирования является предоставление авиационной индустрии и органам безопасности инструмента, способного на раннем этапе предсказывать потенциальное количество жертв в случае инцидента.

2. Моделирование статистических зависимостей

2.1 Формализация и классификация переменных

Были рассмотрены 10 переменных, которые потенциально могут быть полезными в предсказании количества летальных исходов при авиационных инцидентах:

1. X_1 «Aircraft» - качественная переменная, представляет собой множество возможных моделей самолётов.
2. X_2 «Flight phase» - качественная переменная, представляет собой множество стадий полёта, в которых мог находиться самолёт при крушении.
3. X_3 «Flight type» - качественная переменная, множество типов полёта, совершающихся воздушным судном.
4. X_4 «Crash site» - качественная переменная, множество мест, где мог упасть самолёт. Например: поле, лес или аэропорт.
5. X_5 «YOM» - количественная дискретная величина, год выпуска воздушного судна.
6. X_6 «Country» - качественная переменная, множество стран.
7. X_7 «Region» - качественная переменная, множество регионов мира.
8. X_8 «Crew on board» - количественная дискретная переменная, количество персонала на борту самолёта.
9. X_9 «Pass on board» - количественная дискретная переменная, количество пассажиров на борту самолёта.
10. X_{10} «Crash cause» - качественная переменная, множество причин падения самолётов.

Выходной переменной y является «Total fatalities» - количественная дискретная величина, общее количество погибших во время инцидента.

2.2 Проверка гипотезы о нормальном распределении выходной величины

Проверка гипотезы о нормальном распределении была осуществлена с помощью «Правил трёх сигм» и критерия Пирсона.

Правило трёх сигм гласит, что с высокой вероятностью случайная величина не отклонится от своего среднего значения более, чем на 3σ , то есть на 3 среднеквадратических отклонения. Более точно – случайная величина подчинена распределению $N(a, \sigma)$, тогда около

68% ее реализации лежат в интервале $(a - \sigma, a + \sigma)$, около 95% ее реализаций лежат в интервале $(a - 2\sigma, a + 2\sigma)$, а 99.7% ее реализаций лежат в интервале $(a - 3\sigma, a + 3\sigma)$. Применяя данное правило были вычислены значения: 94.65, 97.15, 98.215, что не соответствует правилу трёх сигм и сигнализирует о том, что выходная величина не подчинена закону нормального распределения.

Подтвердим данную теорию с помощью критерия Пирсона [2]. Критерий Хи-квадрат Пирсона используется для проверки гипотезы о соответствии эмпирического распределения предполагаемому. Проверим гипотезу H_0 – величина распределена нормально. Если вычисленный наблюдаемый Хи-квадрат не превышает критическое значение Хи-квадрат, то у нас будет недостаточно оснований отвергнуть H_0 . В данном случае, $\chi_{\text{набл.}}^2$ сильно больше (в 10^{200} раз) $\chi_{\text{крит.}}^2$ следовательно, мы отвергаем гипотезу H_0 о нормальном распределении выходной величины.

В итоге, выходная величина – количество летальных исходов, распределена не нормально, это может ухудшить качество модели, поэтому стоит провести выравнивание для лучших результатов предсказания.

2.3 Корреляционный анализ

Корреляционный анализ используется для определения того, насколько изменения в одной переменной коррелируют с изменениями в другой. Основным инструментом в корреляционном анализе - коэффициент корреляции, чаще всего коэффициент Пирсона.

Коэффициент корреляции принимает значения от -1 до 1 и позволяет оценить характер взаимосвязи между переменными. Значение близкое к 1 указывает на положительную линейную корреляцию, тогда как значение близкое к -1 указывает на отрицательную линейную корреляцию. Коэффициент, близкий к 0, свидетельствует о слабой или отсутствующей линейной связи.

Матрица корреляций показывает коэффициенты корреляции между несколькими переменными.

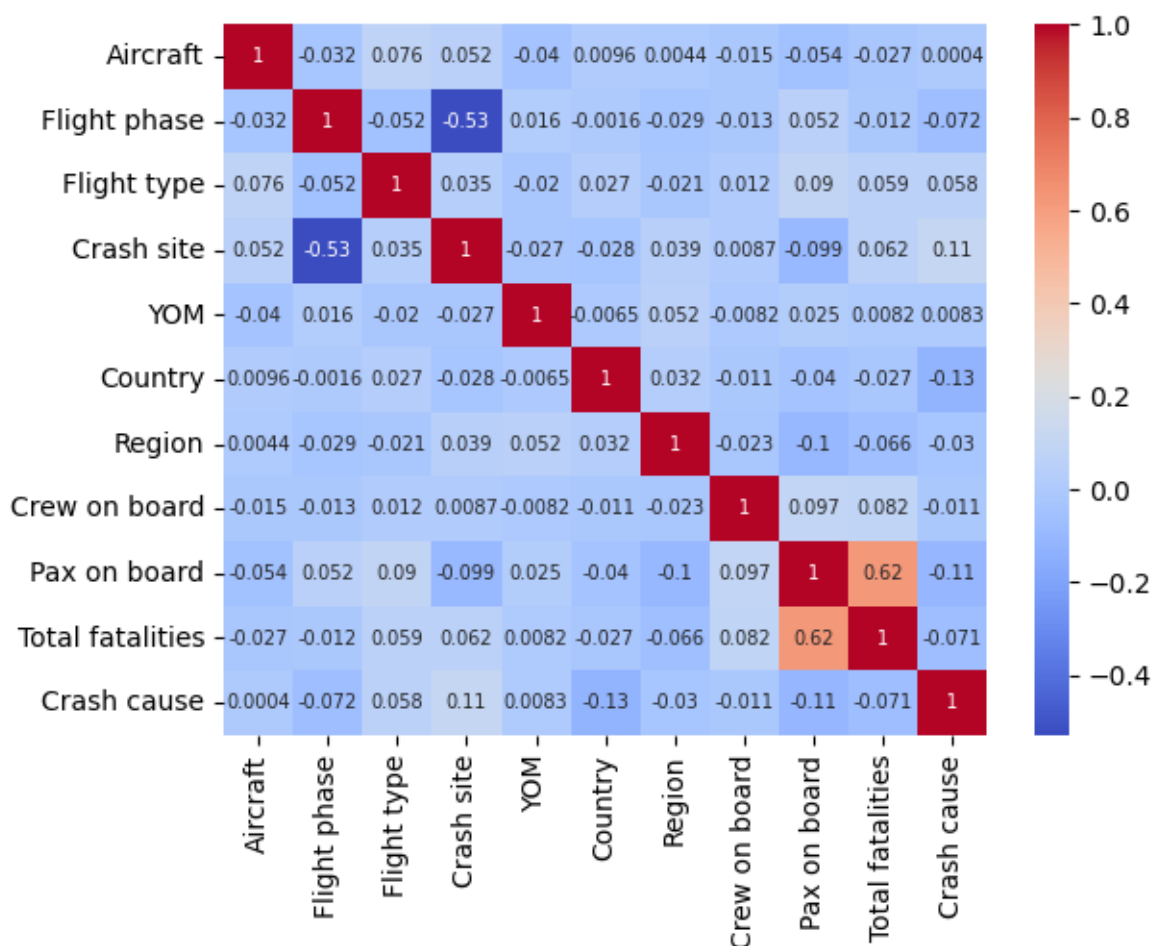


Рис. 1 – Корреляционная матрица

Можно заметить, что многие переменные имеют слабую корреляцию с выходной, а некоторые коррелируют между собой. Корреляция между независимыми переменными называется мультиколлинеарностью, такая связь введет к неопределенности и плохим результатам предсказания.

2.4 Построение регрессионной модели

2.4.1 Структурная идентификация модели

Зависимой переменной является количество смертей. Независимыми переменными являются 10 признаков: модель самолёта, стадия полёта, тип полёта, место падения, год производства самолёта, страна, регион мира, кол-во пассажиров на борту, кол-во персонала на борту, причина происшествя.

Рассмотрим уравнение множественной линейной регрессии $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$, где Y – зависимая переменная, $X_1 \dots X_n$ – независимые переменные, $\beta_0, \beta_1 \dots \beta_n$ –

коэффициенты регрессии, ε – случайная ошибка. Данная функциональная форма отлично подойдет для рассматриваемой задачи.

2.4.2 Параметрическая идентификация модели

В соответствии с методом наименьших квадратов [3], задача заключается в аппроксимации кривой известной функцией. Вычисление параметров уравнения множественной линейной регрессии будет произведено с помощью алгоритма МНК.

Aircraft	0.000049
Flight phase	0.453436
Flight type	-0.001627
Crash site	1.203601
YOM	-0.000233
Country	-0.000014
Region	-0.049325
Crew on board	0.029225
Pax on board	0.435220
Crash cause	-0.160953
intercept	0.000324

Рис. 2 – Результаты МНК

Сразу заметно низкое влияние некоторых коэффициентов. Однако на данной стадии нас интересует лишь полученное уравнение множественной линейной регрессии, которое имеет вид:
$$Y = 0.000324 + 0.000049X_1 + 0.453436X_2 - 0.001627X_3 + 1.203601X_4 - 0.000233X_5 - 0.000014X_6 - 0.049325X_7 + 0.029225X_8 + 0.43522X_9 - 0.160953X_{10}$$

3. Исследование модели

3.1 Анализ статистической значимости уравнения регрессии

Общая сумма квадратов отклонений переменной y от среднего значения \bar{y} может быть разложена на две составляющие: $S_y = S_{\text{факт}} + S_e$, где $S_y = \sum_{i=1}^n (y_i - \bar{y})^2$ - общая сумма квадратов отклонений; $S_{\text{факт}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ - сумма квадратов отклонений, объясненная регрессией; $S_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ - остаточная сумма квадратов отклонений (необъясненная);

Выдвинем гипотезу о равенстве нулю коэффициентов регрессии. В том случае выходная переменная y не зависит от факторов, и вариация y обусловлена только воздействием ошибок: $S_y = S_e$. Противоположным является случай, при котором выходная переменная y функционально зависит от факторов: $S_y = S_{\text{факт}}$.

Для сравнения $S_{\text{факт}}$ и S_e их необходимо разделить на соответствующее число степеней свободы, получив таким образом средний квадрат отклонений на одну степень свободы – дисперсию: $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$, $s_{\text{факт}}^2 = \frac{1}{m} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $s_e^2 = \frac{1}{n-m-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Статистическая значимость уравнения регрессии определяется условием $s_{\text{факт}}^2 > s_e^2$. Задача сводится к проверке нулевой гипотезы $H_0: D_{\text{факт}} = D_e$ при конкурирующей гипотезе $H_1: D_{\text{факт}} > D_e$. Оценка статистической значимости уравнения регрессии выполняется с помощью F-критерия Фишера: $F = \frac{s_{\text{факт}}^2}{s_e^2}$.

Уравнение регрессии является статистически значимым, если:

1. F попадает в критическую область при заданном уровне значимости α , то есть $F > F_{\text{кр}}$
2. Уровень значимости α_F , для которого F является критической точкой (вероятность нулевой гипотезы, Р-значение) меньше заданного уровня значимости α , то есть $\alpha_F < \alpha$.

Для данной модели $F = 1502.28$, $\alpha_F = 0.0014$ при $F_{\text{кр}}, \alpha = 0.05$, что удовлетворяет заданным условиям. Следовательно, текущее уравнение регрессии можно назвать статистически значимым.

3.2 Анализ статистической значимости коэффициентов уравнения регрессии

Для проверки значимости коэффициентов формулируются гипотезы: $H_0: \beta_j = 0$ (коэффициент незначим), $H_1: \beta_j \neq 0$ (коэффициент значим). В качестве критерия выбирается случайная величина T_j , распределенная по закону Стьюдента с $n - m - 1$ степенями свободы: $T_j = \frac{\beta_j}{s_j}$, где β_j – коэффициент уравнения регрессии при факторе x_j , s_j – стандартная ошибка коэффициента β_j . $s_j = s \sqrt{[(C^T C)^{-1}]_{jj}}$, где $[(C^T C)^{-1}]_{jj}$ – j-й диагональный элемент матрицы $(C^T C)^{-1}$, $s = \sqrt{s_e^2}$.

Коэффициент β_j статистически значим, то есть значимо отличается от нуля (принимается гипотеза H_1 на уровне значимости α), если:

1. T_j попадает в критическую область при заданном уровне значимости α , то есть $|T_j| > T_{кр}$;
2. Уровень значимости α_{T_j} , для которого T_j является критической точкой (Р-значение) меньше заданного уровня значимости α : $\alpha_{T_j} < \alpha$

Интервальная оценка для коэффициентов β_j определяется с помощью доверительного интервала $(\beta_j - t_\gamma s_j; \beta_j + t_\gamma s_j)$, где $t_\gamma = t(\alpha, n - m - 1)$.

	coef	std err	t	P> t	[0.025	0.975]
aircraft	4.901e-05	0.000	0.125	0.900	-0.001	0.001
flight_phase	0.4534	0.097	4.684	0.000	0.264	0.643
flight_type	-0.0016	0.012	-0.137	0.891	-0.025	0.022
crash_site	1.2036	0.053	22.924	0.000	1.101	1.307
yom	-0.0002	0.000	-0.700	0.484	-0.001	0.000
country	-1.366e-05	0.001	-0.010	0.992	-0.003	0.003
region	-0.0493	0.050	-0.994	0.320	-0.147	0.048
crew_on_board	0.0292	0.007	3.991	0.000	0.015	0.044
pax_on_board	0.4352	0.004	118.209	0.000	0.428	0.442
crash_cause	-0.1610	0.055	-2.944	0.003	-0.268	-0.054

Рис. 3 – Статистическая значимость коэффициентов регрессии

Работая с уровнем значимости $\alpha = 0.05$, заметны коэффициенты, которые являются статистически незначимыми, то есть они оказывают незначительное влияние на нашу модель. Избавление от таких коэффициентов может привести к лучшим результатам предсказания.

3.3 Исследование мультиколлинеарности факторов

Мультиколлинеарность модели множественной регрессии – наличие высокой взаимной коррелированности между факторами. Последствия мультиколлинеарности:

- Матрица $(C^T C)$ может являться невырожденной, но величина её определителя мала и, как следствие, элементы обратной матрицы становятся очень большими. В результате получаются большие дисперсии коэффициентов;
- Оценки коэффициентов чувствительны к незначительному изменению результатов наблюдений и объема выборки, что делает модель непригодной для анализа и прогнозирования;
- Уменьшаются t-статистики коэффициентов, и оценка их значимости по t-критерию теряет смысл;

Если в матрице парных коэффициентов корреляции факторов пары переменных имеют высокие коэффициенты корреляции, в модели наблюдается мультиколлинеарность. Если же факторы не коррелированы между собой, матрица парных корреляций является единичной матрицей, и ее определитель равен 1. Но если между факторами существует зависимость, то все коэффициенты корреляции равны единице, а определитель равен нулю. Следовательно, чем ближе определитель матрицы парных корреляций к нулю, тем сильнее мультиколлинеарность факторов и наоборот.

Исходя из построенной матрицы парных корреляций (см. рисунок 1), сильная корреляция признаков отсутствует, а значит явление мультиколлинеарности в данной модели не наблюдается. Подтверждение данной гипотезы составляет определитель матрицы равный 0.3865, что является значением, далёким от нуля.

3.4 Применение шагового регрессионного анализа для улучшения модели

Шаговый регрессионный анализ реализуется двумя способами. С помощью добавления факторов и с помощью их удаления. При добавлении определяется фактор, имеющий

наиболее высокий коэффициент корреляции с выходной величиной, а после происходит пошаговое добавление остальных факторов исходя из условия увеличения скорректированного коэффициента детерминации. При удалении факторов берется модель с максимальным числом переменных, на каждом шаге проводится удаление наименее значимого фактора. Изначальный $R_{adj}^2 = 0.39937$.

Шаги при удалении:

1. Удаление «Country» со Р-значением 0.992 приводит к $R_{adj}^2 = 0.39939$
2. Удаление «Aircraft» со Р-значением 0.900 приводит к $R_{adj}^2 = 0.39942$
3. Удаление «Flight type» со Р-значением 0.898 приводит к $R_{adj}^2 = 0.39944$
4. Удаление «YOM» с Р-значением 0.479 приводит к $R_{adj}^2 = 0.3994578$
5. Удаление «Region» с Р-значением 0.161 приводит к $R_{adj}^2 = 0.3994559$

На пятом шаге и далее происходит уменьшение оценки. Следовательно, признаки, которые необходимо использовать: «Flight phase», «Region», «Crash site», «Crew on board», «Pax on board», «Crash cause».

Шаги при добавлении:

1. Добавление «Pax on board» с приводит к $R_{adj}^2 = 0.38304$
2. Добавление «Crew site» приводит к $R_{adj}^2 = 0.39828$
3. Добавление «Flight phase» приводит к $R_{adj}^2 = 0.39884$
4. Добавление «Crew on board» приводит к $R_{adj}^2 = 0.39925$
5. Добавление «Crash cause» приводит к $R_{adj}^2 = 0.3994559$
6. Добавление «Region» приводит к $R_{adj}^2 = 0.3994578$

Дальнейшее добавление признаков приводит к уменьшению R_{adj}^2 . Таким образом, обе методики показали одинаковые результаты.

4. Программная реализация и численное исследование результатов моделирования

4.1 Обоснование выбора и описание программного обеспечения

В процессе работы был использован Python 3 в качестве основного языка программирования, а Jupyter Notebook – в качестве среды разработки и анализа данных. Этот выбор обоснован широкими возможностями, предоставляемыми Python: лаконичным синтаксисом, обширной стандартной библиотекой и активным сообществом разработчиков, что значительно облегчает разработку и поддержку кода.

Для обработки и анализа данных была задействована библиотека Pandas, предоставляющая удобные инструменты для работы с табличными данными. Библиотеки Scikit-learn и Statsmodels использовались для реализации статистических моделей, предоставляя разнообразные алгоритмы для анализа данных и построения моделей.

Matplotlib и Seaborn были использованы для визуализации результатов и изучения структуры данных. Гибкость и функциональность этих библиотек обеспечили создание информативных графиков, способствуя глубокому пониманию данных.

Перечисленные инструменты обеспечили успешное выполнение задач и достижение целей в рамках работы.

4.2 Описание основных модулей программы

Для начала были импортированы все необходимые библиотеки, которые будут использованы на протяжении всей работы.

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.model_selection import cross_val_score
import math
import scipy.stats as ss
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
import matplotlib.pyplot as plt
from scipy.stats import f
from scipy.stats import boxcox
```

Рис. 4 – Импортирование библиотек

Определена функция, которая совершает предобработку датафрейма – отбрасывает ненужные признаки, удаляет данные с неопределенными значениями признаков, изменяет стиль написания названий переменных и преобразует категориальные признаки в числовые. Модель не может работать с неопределенными значениями, поэтому они были отброшены. Изменение имен переменных делает работу с ними в дальнейшем более удобной. Кодирование категориальных признаков в числовые было произведено с помощью Label Encoder, этот тип кодирования используется чаще всего, присваивает число каждому уникальному значению категориального признака.

```
def feature_engineering(df):  
    df = df.drop(["Operator", "Registration", "Survivors", "MSN", "Date", "Time",  
                 "Schedule", "Crew fatalities", "PAX fatalities", "Other fatalities",  
                 "Circumstances", "Flight no.", "Crash location"], axis=1)  
    df = df.dropna()  
    new_names = {name: "_".join(name.lower().split()) for name in df.columns}  
    df = df.rename(columns=new_names)  
    label_encoder = LabelEncoder()  
    df = df.apply(lambda col: label_encoder.fit_transform(col) if col.dtype == 'O' else col)  
    return df  
  
df = feature_engineering(df)
```

Рис. 5 – Функция обработки признаков

Функция, реализующая корреляционный анализ, выводит на экран матрицу парных корреляций и её детерминант.

```
def corr_analysis(df):  
    sns.heatmap(df.corr(), annot=True, cmap="coolwarm", annot_kws={"size": 7})  
    print(f'Детерминант матрицы парных корреляций: {np.linalg.det(df.corr().to_numpy())}')  
  
corr_analysis(df)
```

Рис. 6 – Функция корреляционного анализа

Были реализованы функции для проверки нормальности выходной переменной. Функция проверки правила трёх сигм выводит проценты вхождений в интервалы: одной, двух и трёх сигм. Функция проверки нормальности распределения с помощью критерия Пирсона использует вспомогательную функцию, делящую данные на интервалы.


```

# Проверка на нормальное распределение выходной величины y

# Правило трёх сигм
data = y
mean_data = np.mean(data)
std_data = np.std(data)

def get_interval(k_sigma, mean_value, std_value):
    return [mean_value - k_sigma * std_value, mean_value + k_sigma * std_value]

def get_percent(k, data):
    mean_data = np.mean(data)
    std_data = np.std(data)

    interval = get_interval(k, mean_data, std_data)

    return sum([1 if x >= interval[0] and x <= interval[1] else 0 for x in data]) / len(data) * 100

percent_68 = get_percent(1, data)
percent_95 = get_percent(2, data)
percent_99 = get_percent(3, data)
print(f"Процент вхождений в интервал 1 сигмы: {percent_68}")
print(f"Процент вхождений в интервал 2 сигм: {percent_95}")
print(f"Процент вхождений в интервал 3 сигм: {percent_99}")

```

Рис. 7 – Функция правила трёх сигм

```

def get_interval_distribution(data):
    k = int(np.round(1 + 3.322 * np.log10(len(data))))
    min_age = math.floor(min(data))
    max_age = math.ceil(max(data))
    h = (max_age - min_age) / k

    intervals = [(min_age + i * h, min_age + (i + 1) * h) for i in range(k)]
    distr = {intr:0 for intr in intervals}
    for x in data:
        for intr in intervals:
            if x >= intr[0] and x <= intr[1]:
                distr[intr] += 1
            break
    return (distr, k, h)

def chi_square(data):
    distr, k, h = get_interval_distribution(data)
    sum_xn = 0
    sum_x2n = 0
    # Вычисляем среднее, дисперсию и среднекв. отклонение (x - середина интервала)
    for inter, n in distr.items():
        lower, upper = inter[0], inter[1]
        x = (upper + lower) / 2
        sum_xn += x * n
        sum_x2n += (x ** 2) * n
    mean_value = sum_xn / len(data)
    variance = (sum_x2n / len(data)) - (mean_value ** 2)
    std_value = variance ** 0.5
    # Строим массив теоретических частот
    theor_freqs = []
    freqs = []
    for inter, n in distr.items():
        lower, upper = inter[0], inter[1]
        x = (upper + lower) / 2
        z = (x - mean_value) / std_value
        f_z = math.exp(-(z**2) / 2) / math.sqrt(2 * math.pi)
        theor_n = ((h * len(data)) / std_value) * f_z
        theor_freqs.append(theor_n)
        freqs.append(n)
    # Хи-квадрат наблюдаемое
    chi2_obs = sum([(freqs[i] - theor_freqs[i]) ** 2] / theor_freqs[i] for i in range(len(distr))])
    print(chi2_obs)
    # Хи-квадрат критическое
    chi2_crit = ss.chi2.ppf(1 - 0.05, k - 1)
    print(chi2_crit)
    plt.plot(np.linspace(min(data), max(data), len(data)), data)
    return "Отвергаем H0" if chi2_obs > chi2_crit else "Недостаточно оснований отвергнуть H0"
print("Пусть гипотеза H0 - совокупность распределена нормально")
print("На основании проведенного теста с помощью критерия Пирсона: " + chi_square(data))

```

Рис. 8 – Критерий хи-квадрат Пирсона

В следующем коде используется функция `train_test_split` из библиотеки `Scikit-learn`. Данная функция делит исходные данные на тестовые и тренировочные. На тренировочных данных модель обучается, а на тестовых объективно оценивается качество модели.

```
y = df["total_fatalities"]
X = df.drop(["total_fatalities", "country", "aircraft", "flight_type", "yom"], axis=1)
```

Рис. 9 – Разделение датафрейма на входную и выходную переменные

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

Рис. 10 – Разделение данных на тестовые и тренировочные

Метод наименьших квадратов был реализован с помощью функции `OLS` из библиотеки `Statsmodels`. Данная модель была обучена на тренировочных данных, после чего получен массив `y_pred` – предсказания для тестовых данных. Вспомогательная функция `with_intercept` добавляет новый столбец в матрицу независимых переменных полностью заполненный единицами, это необходимо для вычисления значения `intercept`.

```
def with_intercept(X):
    new_X = X.copy()
    new_X["intercept"] = 1
    return new_X

model = sm.OLS(y_train, with_intercept(X_train)).fit()
y_pred = model.predict(with_intercept(X_test))
y_pred = list(map(round, y_pred))
```

Рис. 11 – Программная реализация МНК

Для оценки качества модели была написана функция, которая выводит на экран абсолютную ошибку среднего, среднеквадратичную ошибку, коэффициент детерминации и его исправленную версию.

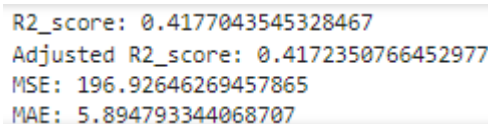
```
def print_scores(y_test, y_pred, k):
    n = len(y_test)
    r2 = r2_score(y_test, y_pred)
    adj_r2 = 1 - (1 - r2) * (n - 1) / (n - k - 1)
    mae = mean_absolute_error(y_test, y_pred)
    mse = mean_squared_error(y_test, y_pred)
    print(f"R2_score: {r2}")
    print(f"Adjusted R2_score: {adj_r2}")
    print(f"MSE: {mse}")
    print(f"MAE: {mae}")
    print_scores(y_test, y_pred, len(X.columns))
```

Рис. 12 – Функция вывода оценок

4.3 Численное исследование результатов моделирования

Оценка модели была произведена по 4 характеристикам, оценивающим качество предсказаний и модели в целом:

- MSE (Mean Squared Error) – средняя квадратичная ошибка. Измерение среднего квадрата разности между предсказанными и фактическими значениями.
- MAE (Mean Absolute Error) – средняя абсолютная ошибка. Измерение среднего значения абсолютных разностей между предсказанными и фактическими значениями.
- R^2 score – коэффициент детерминации. Измеряет долю дисперсии зависимой переменной, которая может быть объяснена моделью. Принимает значения от 0 до 1, где 1 означает идеальное предсказание.
- Adjusted R^2 score – скорректированный коэффициент детерминации, учитывающий количество предикторов в модели и корректирующий R^2 score в случае наличия избыточных предикторов.



```
R2_score: 0.4177043545328467
Adjusted R2_score: 0.4172350766452977
MSE: 196.92646269457865
MAE: 5.894793344068707
```

Рис. 13 – Характеристики модели

Коэффициент детерминации достаточно низок, это свидетельствует о низком качестве предсказаний. MSE принимает высокое значение. MAE примерно равно 6, в контексте рассматриваемой задачи это значит, что предсказанной значение потенциального количества смертей может отличаться от истинного до ± 6 .

Все характеристики свидетельствуют о том, что модель плохо выполняет свою задачу, поэтому следует произвести улучшение модели, чтобы добиться приемлемого качества предсказания.

4.4 Улучшение качества предсказания с помощью методов машинного обучения

Существуют разные методики для улучшения качества предсказания: различные модели машинного обучения, подбор гиперпараметров, тщательная предобработка данных, написание собственной нейронной сети. Применение приведенных методов потенциально

может улучшить модель до неплохого уровня, однако это выходит за рамки выполняемой работы. Для демонстрации потенциала данного исследования было принято решение использовать в качестве модели-регрессора не множественную линейную регрессию, а алгоритм машинного обучения известный как метод случайного леса [4] без подобранных гиперпараметров.

```
R2_score: 0.6011612571432967  
Adjusted R2_score: 0.6008398290093625  
MSE: 134.883204825123  
MAE: 3.8520942387238177
```

Рис. 14 – Оценки улучшенной модели

По сравнению с предыдущей моделью качество предсказания значительно улучшилось. При дальнейшей реализации методик улучшения модель действительно может использовать в различных прикладных задачах.

Выводы

В ходе проведенного исследования были выявлены статистические зависимости между количеством смертей при авиационных инцидентах и рядом факторов. Анализ характеристик объекта исследования позволил определить ключевые переменные, оказывающие влияние на исследуемый показатель. В результате формализации и классификации переменных была проверена гипотеза о нормальном распределении выходной величины.

Корреляционный анализ подтвердил наличие статистически значимых связей между переменными, а построение регрессионной модели дало возможность выявить структурные и параметрические характеристики влияющих факторов.

Исследование модели подтвердило статистическую значимость уравнения регрессии, а также позволило провести анализ статистической значимости коэффициентов уравнения. Применение шагового регрессионного анализа способствовало улучшению модели, оптимизации коэффициентов и исключению мультиколлинеарности факторов.

В ходе программной реализации и численного исследования результатов моделирования было обосновано выбор программного обеспечения, представлено описание основных модулей программы и проведено численное исследование результатов. Улучшение качества предсказания с привлечением методов машинного обучения дополнительно подчеркнуло релевантность и практическую применимость разработанной модели.

Приложение А

Список использованных источников

- 1 Архив инцидентов // Бюро архивов авиационных происшествий. – URL: <https://www.baaa-acro.com/index.php/crash-archives/> (дата обращения 16.12.2023)
- 2 Бослаф С. Статистка для всех. М.: ДМК Пресс, 2015
- 3 Линник Ю.В. Метод наименьших квадратов и основы теории обработки наблюдений. М.: Государственное издательство физико-математической литературы, 1962
- 4 Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015
- 5 Маккинни У. Python и анализ данных. М.: Nobel Press, 2020

Приложение Б

Программная реализация:

URL: <https://github.com/dkshi/aviation>