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.

**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. **Описание объекта исследования**

Собранная информация с интернет-архива авиационных инцидентов [], 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 длинной 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 , 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 , где N – количество значений. Дисперсия примерно равна 279.3312.

Среднее квадратическое отклонение, вычисленное по формуле 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: . 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 . Если число вариант четное, то , 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 индустрии 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. «Aircraft» - качественная переменная, представляет собой множество возможных моделей самолётов.
2. «Flight phase» - качественная переменная, представляет собой множество стадий полёта, в которых мог находится самолёт при крушении.
3. «Flight type» - качественная переменная, множество типов полёта, совершающихся воздушным судном.
4. «Crash site» - качественная переменная, множество мест, где мог упасть самолёт. Например: поле, лес или аэропорт.
5. «YOM» - количественная дискретная величина, год выпуска воздушного судна.
6. «Country» - качественная переменная, множество стран.
7. «Region» - качественная переменная, множество регионов мира.
8. «Crew on board» - количественная дискретная переменная, количество персонала на борту самолёта.
9. «Pax on board» - количественная дискретная переменная, количество пассажиров на борту самолёта.
10. «Crash cause» - качественная переменная, множество причин падения самолётов.

Выходной переменной 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 «Правила трёх сигм» и критерия Пирсона.

Правило трёх сигм гласит, что с высокой вероятность случайная величина не отклонится от своего среднего значения более, чем на 3, то есть на 3 среднеквадратических отклонения. Более точно – случайная величина подчинена распределению , тогда около 68% ее реализации лежат в интервале , около 95% ее реализаций лежат в интервале , а 99.7% ее реализаций лежат в интервале 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 []. Критерий Хи-квадрат Пирсона используется для проверки гипотезы о соответствии эмпирического распределения предполагаемому. Проверим гипотезу – величина распределена нормально. Если вычисленный наблюдаемый Хи-квадрат не превышает критическое значение Хи-квадрат, то у нас будет недостаточно оснований отвергнуть . В данном случае, сильно больше (в раз) следовательно, мы отвергаем гипотезу 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. Основной инструмент 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.

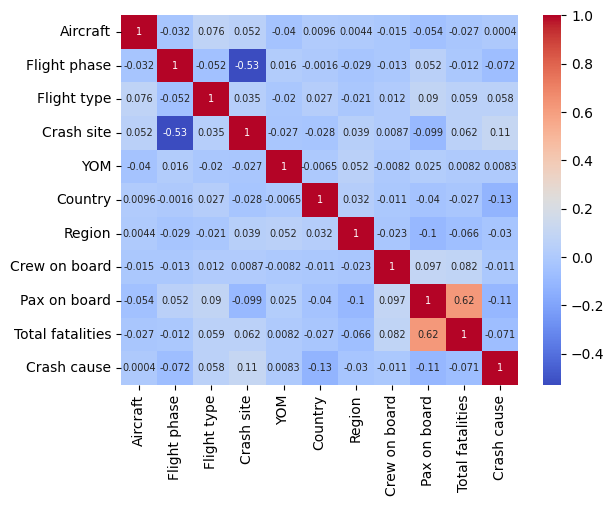


Рис. 1 – Корреляционная матрица

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 происшествия.

Рассмотрим уравнение множественной линейной регрессии , где Y – зависимая переменная, -независимые переменные, - коэффициенты регрессии, – случайная ошибка. Данная функциональная форма отлично подойдет для рассматриваемой задачи.

**2.4.2 Параметрическая идентификация модели**

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

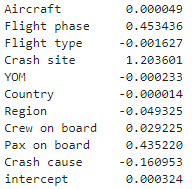


Рис. 2 – Результаты МНК

Сразу заметно низкое влияние некоторых коэффициентов. Однако на данной стадии нас интересует лишь полученное уравнение множественной линейной регрессии, которое имеет вид:

**3. Исследование модели**

**3.1 Анализ статистической значимости уравнения регрессии**

Общая сумма квадратов отклонений переменной от среднего значения может быть разложена на две составляющие: , где - общая сумма квадратов отклонений; - сумма квадратов отклонений, объясненная регрессией; – остаточная сумма квадратов отклонений (необъясненная);

Выдвинем гипотезу о равенстве нулю коэффициентов регрессии. В том случае выходная переменная не зависит от факторов, и вариация обусловлена только воздействием ошибок: . Противоположным является случай, при котором выходная переменная функционально зависит от факторов: .

Для сравнения и их необходимо разделить на соответствующее число степеней свободы, получив таким образом средний квадрат отклонений на одну степень свободы – дисперсию: , ,

Статистическая значимость уравнения регрессии определяется условием . Задача сводится к проверке нулевой гипотезы при конкурирующей гипотезе . Оценка статистической значимости уравнения регрессии выполняется с помощью F-критерия Фишера: .

Уравнение регрессии является статистически значимым, если:

1. F попадает в критическую область при заданном уровне значимости , то есть
2. Уровень значимости , для которого является критической точкой (вероятность нулевой гипотезы, P-значение) меньше заданного уровня значимости , то есть .

Для данной модели при , 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: (коэффициент незначим), (коэффициент значим). В качестве критерия выбирается случайная величина , распределенная по закону Стьюдента с степенями свободы: , где – коэффициент уравнения регрессии при факторе , – стандартная ошибка коэффициента . , где – j-й диагональный элемент матрицы , .

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

1. попадает в критическую область при заданном уровне значимости , то есть ;
2. Уровень значимости , для которого является критической точкой (P-значение) меньше заданного уровня значимости

Интервальная оценка для коэффициентов определяется с помощью доверительного интервала , где .

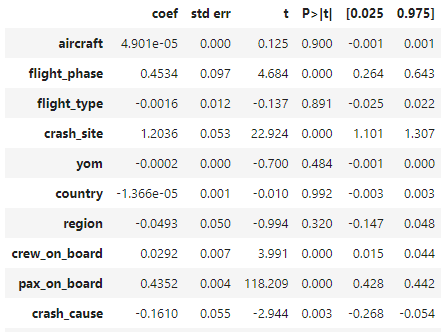


Рис. 3 – Статистическая значимость коэффициентов регрессии

Работая с уровнем значимости , заметны коэффициенты, которые являются статистически незначимыми, то есть они оказывают незначительно 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 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 наблюдает 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 , 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 .

Шаги при удалении:

1. Удаление «Country» со P-значением 0.992 приводит к
2. Удаление «Aircraft» со P-значением 0.900 приводит к
3. Удаление «Flight type» со P-значением 0.898 приводит к
4. Удаление «YOM» с P-значением 0.479 приводит к
5. Удаление «Region» с P-значением 0.161 приводит к

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
2. Добавление «Crew on board» приводит к
3. Добавление «Crash cause» приводит к
4. Добавление «Region» приводит к
5. Добавление «Crash site» приводит к

Дальнейшее добавление признаков приводит к уменьшению . Таким образом, лучший измененный коэффициент детерминации показала методика удаления признаков.