ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»»

Высшая школа бизнеса

Департамент бизнес-информатики

Отчёт по проделанной работе

по теме

«Моделирование удовлетворенности пассажиров авиаперелетом с использованием предиктивной аналитики на базе IBM SPSS Modeler»

Работу выполнил студент

2 курса группы МБИ-202

Андросов А.К.

Научный руководитель

к.т.н., доцент

Бекларян А.Л.

Москва, 2021

СОДЕРЖАНИЕ

[1. ПОСТАНОВКА ЗАДАЧИ И ОПИСАНИЕ ДАННЫХ 3](#_Toc89167467)

[2. МОДИФИКАЦИЯ ДАННЫХ В IBM SPSS MODELER 6](#_Toc89167468)

[3. ПЕРВИЧНЫЙ АНАЛИЗ ВЕРХНЕГО УРОВНЯ В СИСТЕМЕ POWER BI 13](#_Toc89167469)

[4. АНАЛИЗ И ПОДГОТОВКА ДАННЫХ В IBM SPSS MODELER 27](#_Toc89167470)

[5. МОДЕЛИРОВАНИЕ И ОЦЕНКА КАЧЕСТВА МОДЕЛИ 38](#_Toc89167471)

[6. ПРОВЕРКА МОДЕЛЕЙ НА ТЕСТОВЫХ ДАННЫХ 48](#_Toc89167472)

[7. ВЫВОДЫ ПО РАБОТЕ 53](#_Toc89167473)

# 1. ПОСТАНОВКА ЗАДАЧИ И ОПИСАНИЕ ДАННЫХ

Компания «Air» — крупная частная авиакомпания с современным парком воздушных судов на рынке авиаперевозок.

«Air» входит в десятку лучших авиакомпаний Восточной Европы в авторитетном международном рейтинге Skytrax. Один из самых экологичных авиаперевозчиков мира, по оценке экологической организации Atmosfair. Авиакомпания занимает 30 место в мировом рейтинге.

Целью работы является разработка модели прогнозирования удовлетворенности пассажиров авиаперелетом.

Также выявление факторов и устранение недостатков по выявленным факторам, наиболее сильно влияющих на удовлетворенность пассажиров помогут авиакомпании улучшить сервисы для пассажиров, повысить лояльность пассажиров, увеличить количество перелетов разными классами.

В качестве данных был использован датасет «Airline Passenger Satisfaction», взятый с сайта kaggle.com:

Ссылка на датасет: <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction?select=train.csv>).

Источник представляет собой два набора данных – тренировочный и тестовый. В тренировочном наборе данных 103904 записи, в тестовом наборе 25977 записи.

Фрагмент данных представлен на рисунке 1.1.

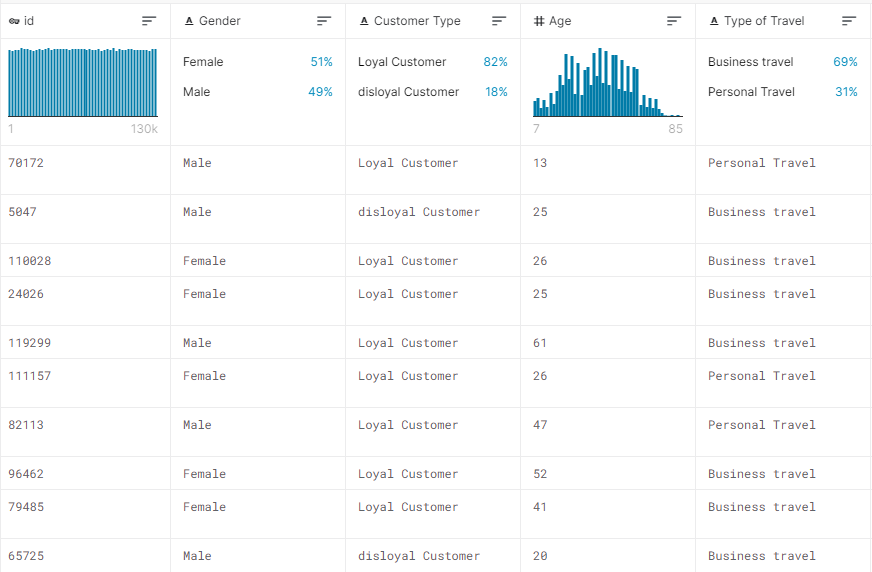


Рисунок 1.1 – Фрагмент данных «Airline Passenger Satisfaction»

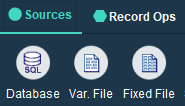
В таблице 1.1 представлены атрибуты данных.

|  |  |
| --- | --- |
| Атрибут | Описание атрибута |
| field1 | Порядковый номер записи. |
| id | Идентификатор пассажира. |
| Gender | Пол пассажира: женский, мужской. |
| Customer Type | Тип пассажира: постоянный клиент, непостоянный клиент. |
| Age | Возраст пассажира. |
| Type of Travel | Цель полета пассажира (личная, деловая). |
| Class | Класс в самолете (Бизнес, Эко, Эко Плюс). |
| Flight distance | Расстояние полета этого путешествия. |
| Inflight wifi service | Уровень удовлетворенности услугой Wi-Fi на борту, где 0 - не приемлемо; 5 – замечательно.  (\*данная шкала будет использована в последующих атрибутах уровня удовлетворенности) |
| Departure/Arrival time convenient | Уровень удовлетворенности удобным временем отправления / прибытия. |
| Ease of Online booking | Уровень удовлетворенности легкостью онлайн-бронирования. |
| Gate location | Уровень удовлетворенности местоположением выхода на посадку. |
| Food and drink | Уровень удовлетворенности едой и напитками. |
| Online boarding | Уровень удовлетворенности онлайн-посадкой. |
| Seat comfort | Уровень удовлетворенности комфортом сиденья. |
| Inflight entertainment | Уровень удовлетворенности развлечениями в полете. |
| On-board service | Уровень удовлетворенности бортовым обслуживанием. |
| Leg room service | Уровень удовлетворенности местом для ног. |
| Baggage handling | Уровень удовлетворенности обработкой багажа. |
| Check-in service | Уровень удовлетворенности службой регистрации. |
| Inflight service | Уровень удовлетворенности обслуживанием в полете. |
| Cleanliness | Уровень удовлетворенности чистотой в самолете. |
| Departure Delay in Minutes | Задержка в минутах при отправлении. |
| Arrival Delay in Minutes | Задержка в минутах при прибытии. |
| Satisfaction | Удовлетворенность авиакомпанией (удовлетворён, нейтрален или неудовлетворён). |

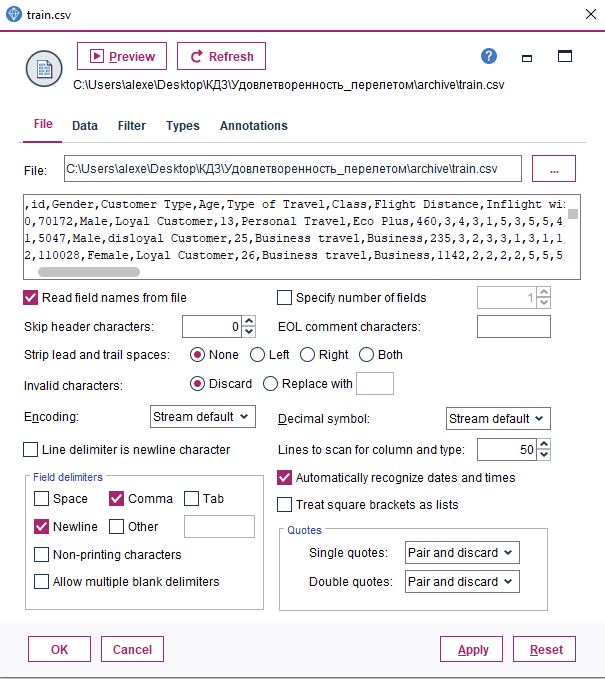
Таблица 1.1 – Атрибуты данных

# 2. МОДИФИКАЦИЯ ДАННЫХ В IBM SPSS MODELER

Для импортирования файла с расширением «.csv» в IBM SPSS Modeler необходимо использовать ноду «Var. File» из палитры «Sources» (рисунок 2.1).

  
Рисунок 2.1 – Нода «Var. File»

В настройках Ноды «Var. File» необходимо указать путь к файлу (рисунок 2.2).

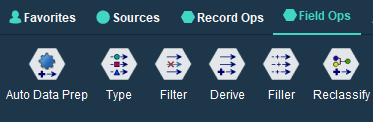
  
Рисунок 2.2 – Настройки для ноды «Var. File»

Затем необходимо определить тип каждого поля в соответствии с таблицей 2.1.

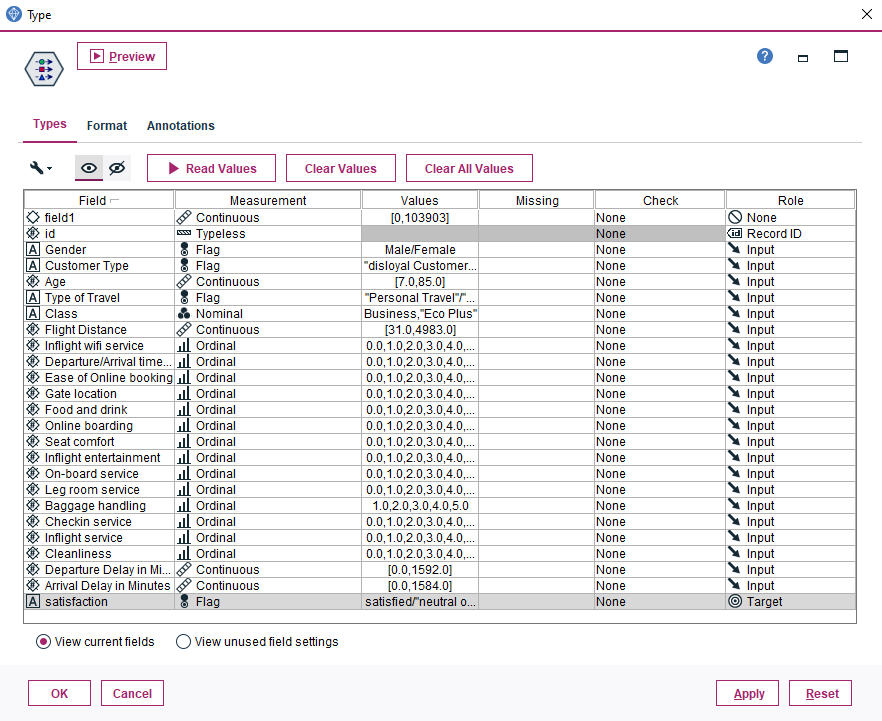
|  |  |  |
| --- | --- | --- |
| Атрибут | Описание атрибута | Тип поля |
| field1 | Порядковый номер записи. | Continuous (Непрерывный).  *Не будет учувствовать в анализе.* |
| id | Идентификатор пассажира. | Typeless (Тип отсутствует) |
| Gender | Пол пассажира: женский, мужской. | Flag (Флаговый) |
| Customer Type | Тип пассажира: постоянный клиент, непостоянный клиент. | Flag (Флаговый) |
| Age | Возраст пассажира. | Continuous (Непрерывный) |
| Type of Travel | Цель полета пассажира (личная, деловая). | Flag (Флаговый) |
| Class | Класс в самолете (Бизнес, Эко, Эко Плюс). | Nominal (Номинативный) |
| Flight distance | Расстояние полета этого путешествия. | Continuous (Непрерывный) |
| Inflight wifi service | Уровень удовлетворенности услугой Wi-Fi на борту, где 0 - не приемлемо; 5 – замечательно.  (\*данная шкала будет использована в последующих атрибутах уровня удовлетворенности, кроме Baggage handling 1-5) | Ordinal (Упорядоченный) |
| Departure/Arrival time convenient | Уровень удовлетворенности удобным временем отправления / прибытия. | Ordinal (Упорядоченный) |
| Ease of Online booking | Уровень удовлетворенности легкостью онлайн-бронирования. | Ordinal (Упорядоченный) |
| Gate location | Уровень удовлетворенности местоположением выхода на посадку. | Ordinal (Упорядоченный) |
| Food and drink | Уровень удовлетворенности едой и напитками. | Ordinal (Упорядоченный) |
| Online boarding | Уровень удовлетворенности онлайн-посадкой. | Ordinal (Упорядоченный) |
| Seat comfort | Уровень удовлетворенности комфортом сиденья. | Ordinal (Упорядоченный) |
| Inflight entertainment | Уровень удовлетворенности развлечениями в полете. | Ordinal (Упорядоченный) |
| On-board service | Уровень удовлетворенности бортовым обслуживанием. | Ordinal (Упорядоченный) |
| Leg room service | Уровень удовлетворенности местом для ног. | Ordinal (Упорядоченный) |
| Baggage handling | Уровень удовлетворенности обработкой багажа. | Ordinal (Упорядоченный) |
| Check-in service | Уровень удовлетворенности службой регистрации. | Ordinal (Упорядоченный) |
| Inflight service | Уровень удовлетворенности обслуживанием в полете. | Ordinal (Упорядоченный) |
| Cleanliness | Уровень удовлетворенности чистотой в самолете. | Ordinal (Упорядоченный) |
| Departure Delay in Minutes | Задержка в минутах при отправлении. | Continuous (Непрерывный) |
| Arrival Delay in Minutes | Задержка в минутах при прибытии. | Continuous (Непрерывный) |
| Satisfaction | Удовлетворенность авиакомпанией (удовлетворен, нейтрален или неудовлетворен). | Flag (Флаговый) |

Таблица 2.1 – Тип полей.

Для того чтобы назначить тип каждому полю, добавляем в поток ноду «Type» из палитры «Field Ops» (рисунок 2.3).

  
Рисунок 2.3 – Нода «Type»

В настройках ноды «Type» указываем типы полей и нажимаем кнопку «Read values» (рисунок 2.4). Также указываем целевую переменную «satisfaction» и установить значение «None» для поля «field1».

Рисунок 2.4 – Настройка типов полей

Далее с помощью ноды «Derive» добавим новое вычислимое поле (рисунок 2.5).



Рисунок 2.5 – Нода «Derive»

Напишем формулу для нового поля «Delay», которое будет означать общую задержку рейса (рисунок 2.6).

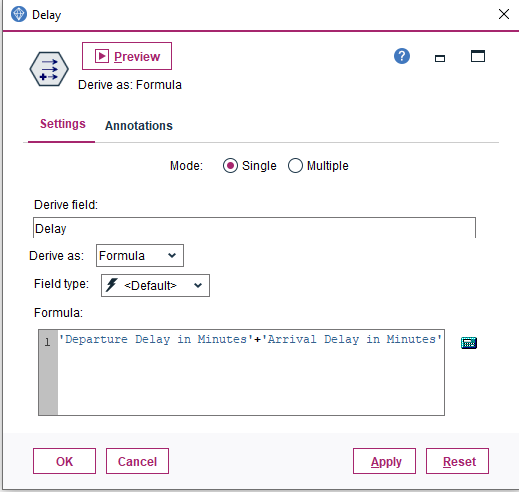


Рисунок 2.6 – Настройки ноды «Derive»

Аналогичным образом добавим еще одно вычислимое поле, такое как общая средняя оценка за все сервиса авиакомпании во время перелета (рисунок 2.7).

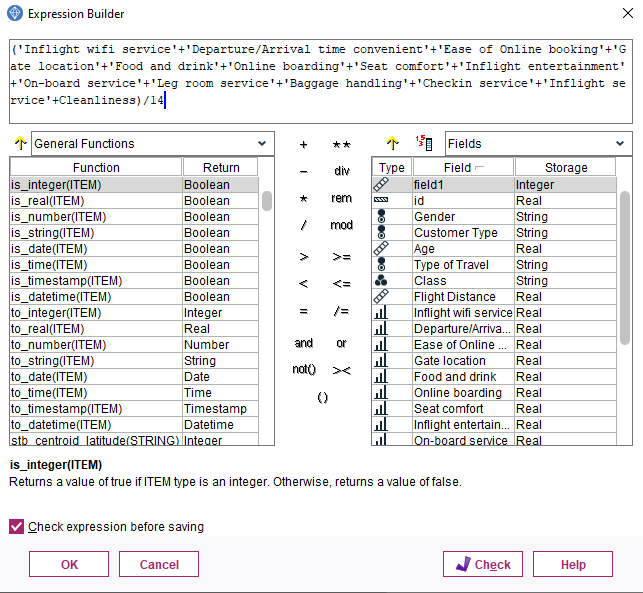


Рисунок 2.7 – Настройки ноды «Derive» для общей оценки сервисов

Определим типы для новых переменных (рисунок 2.8).

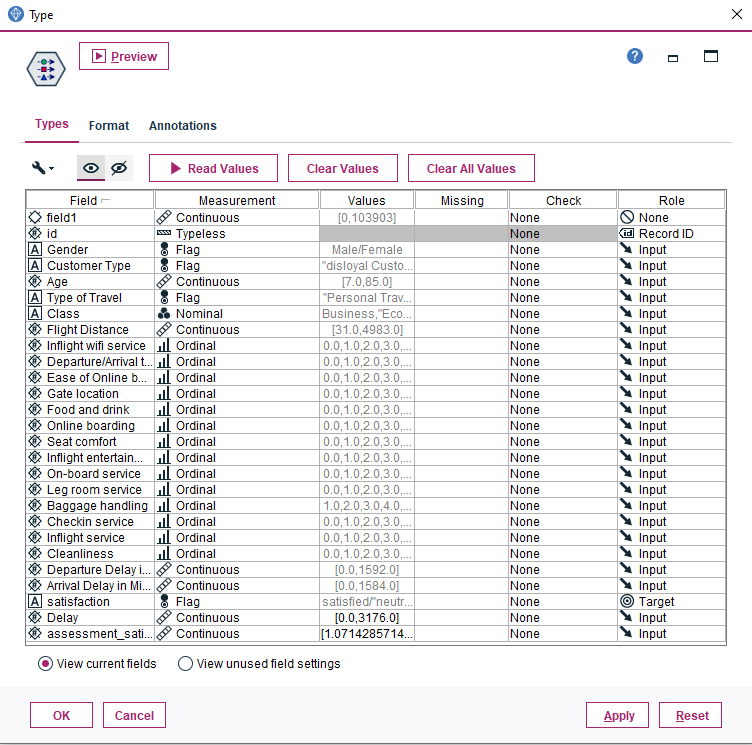


Рисунок 2.8 – Определение типов новых переменных

Далее выгрузим данные в формате «Excel» для дальнейшего анализа в MS Power BI (рисунок 2.9).

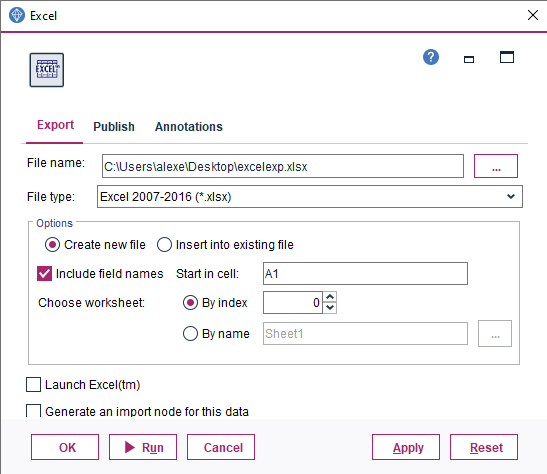


Рисунок 2.9 – Выгрузка данных в формате «Excel»

Аналогичные действия проведем и для тестовой выборки.

# 3. ПЕРВИЧНЫЙ АНАЛИЗ ВЕРХНЕГО УРОВНЯ В СИСТЕМЕ POWER BI

Дополнительную модификацию и верхнеуровневый анализ будем выполнять в Microsoft Power BI. Запускаем Microsoft Power BI, далее выбираем «Get data» 🡪 «Text/CSV» (предварительно были сохранены данные в формате .csv) и через обзор находим необходимую папку с данными и подгружаем данные в формате .csv.

Предварительный просмотр загруженных данных представлен на рисунке 3.1.

## 

Рисунок 3.1 – Предварительный просмотр данных в Power BI

На данном шаге оставляем настройки по-умолчанию, затем переходим в пункт «Transform Data» и проводим следующие действия:

1. Удаляем столбец «Порядковый номер записи».
2. Меняем тип данных у столбца «id» с «Whole number» на «Text» (чтобы Power BI не стал суммировать идентификаторы пассажиров).
3. У столбца «Arrival Delay in Minutes» поменяем тип данных с «Text» на «Whole number».
4. Остальные типы данных оставляем по умолчанию.

Далее правой кнопкой мыши по текущей таблице создадим дубликаты этой таблицы для дальнейшего анализа и операции группировки данных.

Рассчитаем среднюю оценку авиакомпании, выставленных удовлетворенными и неудовлетворенными пассажирами по всем уровням – уровень удовлетворенности едой и напитками, уровень удовлетворенности комфортом сидений, развлечениями в полете и т.д., для этого воспользуемся операцией «Group By», настройки команды приведены на рисунке 3.2.

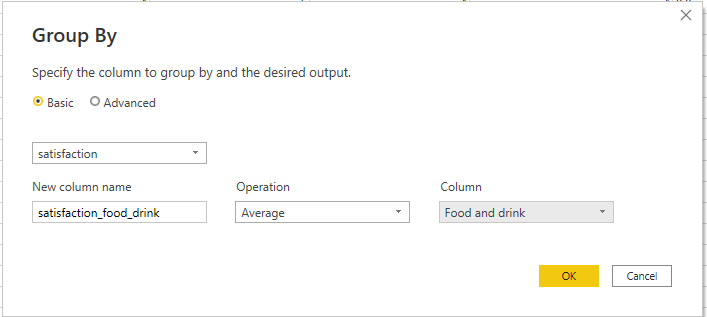


Рисунок 3.2 – Настройки операции «Group By»

Группировку данных будем осуществлять по полю «Satisfaction», новую таблицу назовем «satisfaction\_food\_drink», будем считать операцию «Average» (средняя оценка по еде и напиткам) по колонке «Food and drink».

Результат операции «Group By» представлен на рисунке 3.3.

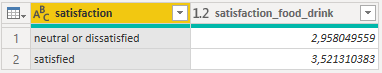


Рисунок 3.3 – Результат операции «Group By» по колонке **«**Food and drink**»**

Далее аналогично по вышеуказанной инструкции определим средние оценки для остальных полей (всего таких таблиц будет 14).

Таблицы со средними оценками услуг приведены на рисунке 3.4.

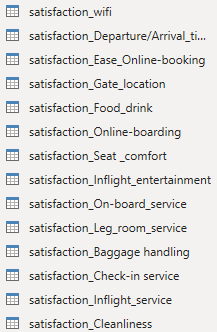


Рисунок 3.4 – Таблицы со средними оценками услуг авиакомпании

В случае если будет необходимо выполнить операцию «Group By» снова, можно будет вернуться к трансформированию данных.

Для того чтобы определить какие оценки по перелету по всем сервисам поставили пассажиры авиакомпании выполним операцию «Group By» снова по «satisfaction». Результат представлен на рисунке 3.5.

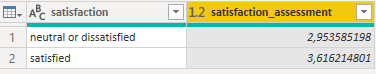


Рисунок 3.5 – Оценка авиакомпании в целом по всем сервисам

Чтобы построить визуализацию процента удовлетворенных пассажиров в зависимости от класса перелета и других категорий необходимо выполнить по две операции группировки данных на каждую категорию. Одна группировка простая, чтобы найти общее количество, другая сложная, чтобы определить количество удовлетворенных пассажиров по нужной категории. Настройки сложной группировки представлены на рисунке 3.6.

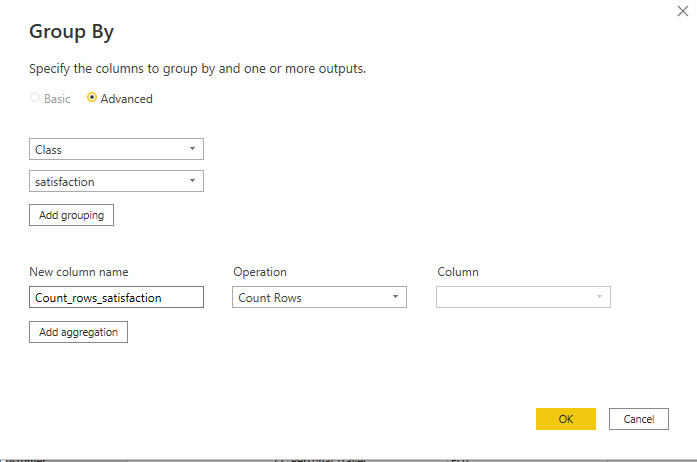


Рисунок 3.6 – Настройка сложной группировки удовлетворенности пассажиров по классу перелета

Затем удалим строки неудовлетворенных пассажиров и в дальнейшем добавим столбец с общим количеством перелетом по каждому классу. Далее создадим новый столбец с соотношением удовлетворенных пассажиров к общему количеству перелетов по каждому классу.

Аналогично выполним по две группировки для типа путешествия и типу пассажира.

Далее загружаем модифицированные данные в Power BI, нажав «Close & Apply».

Следующим шагом будет создание мер через операцию «New measure», посчитаем среднюю оценку авиакомпании по всем пассажирам по сервису «Inflight wifi service».

Формула расчета меры представлена на рисунке 3.7.



Рисунок 3.7 – Формула расчета меры «Inflight wifi service»

Создадим ещё две меры – подсчет общего количества задержек при отправлении и прибытии. Перед созданием этих мер введем два дополнительных столбца для промежуточных вычислений.

В первый столбец по формуле запишем 1, если была задержка при отправлении. Формула расчета общего количества задержек в датасете при отправлении представлена на рисунке 3.8.



Рисунок 3.8 – Формула расчета общего количества задержек при отправлении

Аналогичным образом напишем формулу для расчета общего числа задержек при прибытии. Формула представлена на рисунке 3.9.

 Рисунок 3.9 – Формула расчета общего количества задержек при прибытии

Создадим меру, написав формулу, представленную на рисунке 3.10. С помощью данной формулы, просуммировав количество единиц из столбца «Total\_Delay\_Departure» узнаем общее количество задержек рейсов при отправлении.



Рисунок 3.10 – Формула расчета задержек при отправлении

Далее аналогичным образом подсчитаем количество задержек при прибытии. Формула представлена на рисунке 3.11.



Рисунок 3.11 – Формула расчета задержек при прибытии

Приступим к выполнению первичного анализа и визуализации данных. Для этого сначала рассмотрим визуализацию средних оценок сервисов авиакомпании. Необходимо перейти в раздел «Report» и выбрать следующую визуализацию – «Clustered column chart», виды визуализации представлены на рисунке 3.12.

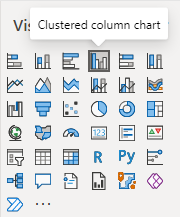


Рисунок 3.12 – Виды визуализации.

Далее проведем настройку визуализации для этого необходимо нажать на саму визуализацию и перейти во вкладку «Fields» и перенести соответствующие ранее созданные меры (средняя оценка по сервису) в поле «Values», исследовать оценки будем по полю «Class» (Business, Eco, Eco plus). Настройки визуализации представлены на рисунке 3.13.

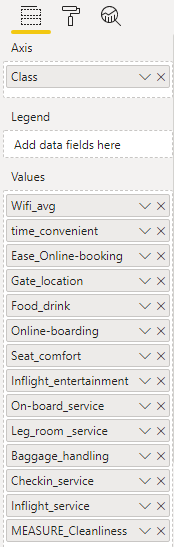


Рисунок 3.13 – Настройки визуализации представлены «Clustered column chart»

Результатом будет визуализация, где можно будет наблюдать средние оценки за сервисы авиакомпании по классу пассажиров. Визуализация «Clustered column chart» представлена на рисунке 3.14.

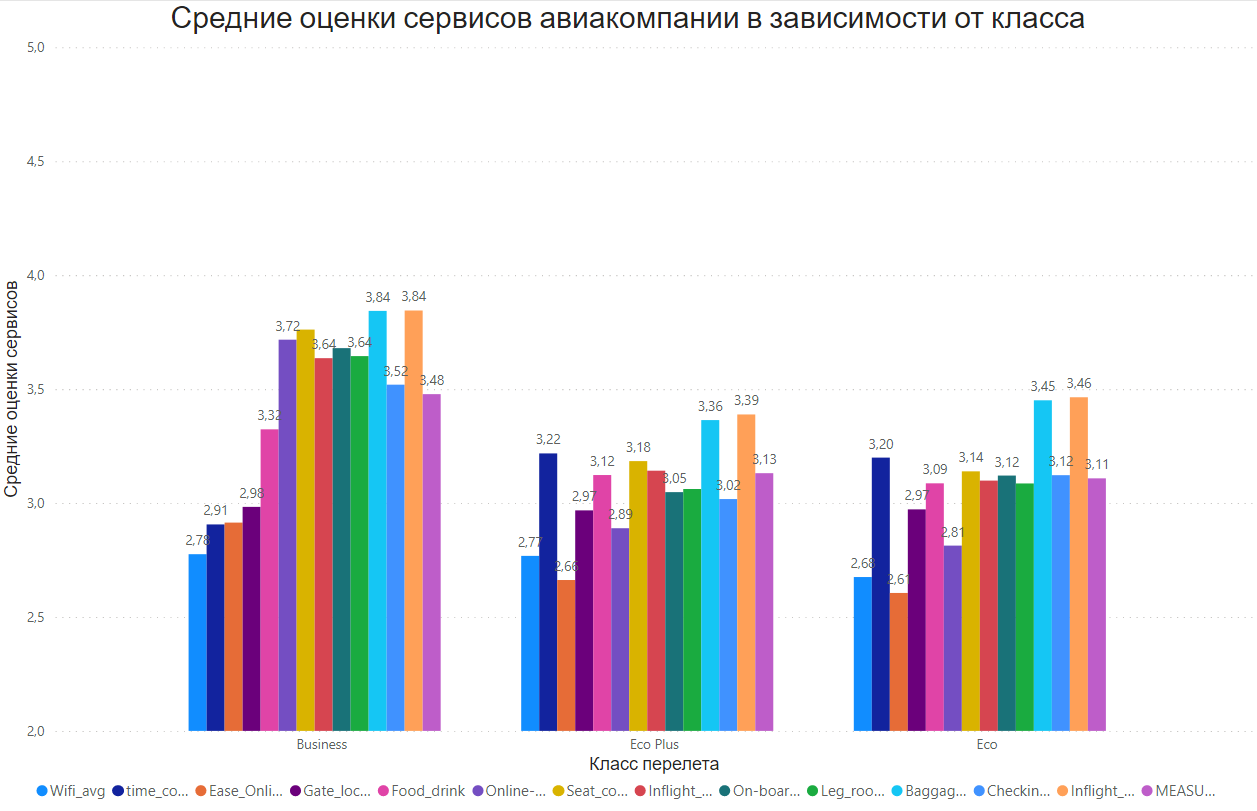


Рисунок 3.14 – Результат визуализации «Clustered column chart»

В результате анализа видно, что среди представленных классов перелета средние оценки по предоставлению вай-фай услуг соизмеримы и пассажиры всех класс неудовлетворенны данным сервисом. Также стоит отметить низкие средние оценки легкости онлайн-бронирования, возможно, авиакомпании следует уделить этому внимание и выяснить причину. Возможно, принять меры по введению инструкций по бронированию, устранения каких либо ошибок, недочетов, четкой структуризации этапов бронирования билетов.

Также для визуализации «Clustered column chart» была проведена настройка формата данных, во вкладке «Format» были включены метки данных с помощью включения «тумблера» поля «Display data label options», а также легенда была перенесена вниз по центру. В данной вкладке для эстетики могут быть настроены и другие элементы, такие как выбор цвета, размера и стиля шрифта и так далее. Настройки «Format» приведены на рисунке 3.15.

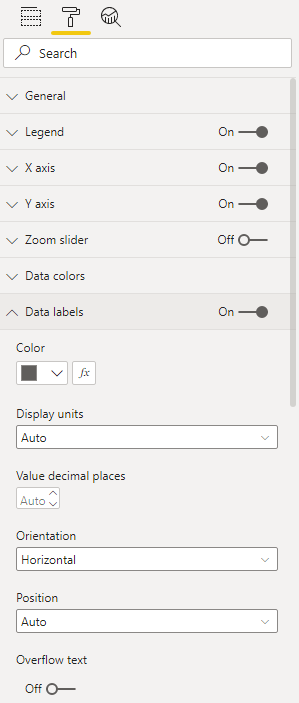


Рисунок 3.15 – Настройка формата визуализации «Clustered column chart»

Также на данный график можно поместить фильтры, для этого необходимо выбрать «Slicer» из видов визуализации, которая представлена на рисунке 3.11. С помощью этого можно будет анализировать оценки пассажиров сервисов авиакомпании в зависимости от заданных фильтров. Примеры фильтров по полу, категории путешествия, типа пассажира и его удовлетворенности перелетом представлены на рисунке 3.16.

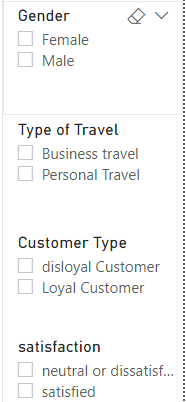


Рисунок 3.16 – «Slicer» по категориям

На рисунке 3.17 представлен пример визуализации оценок нелояльных неудовлетворенных перелетом пассажиров мужского пола, которые летали бизнес-классом по деловым делам. Также можно использовать фильтры по-отдельности.

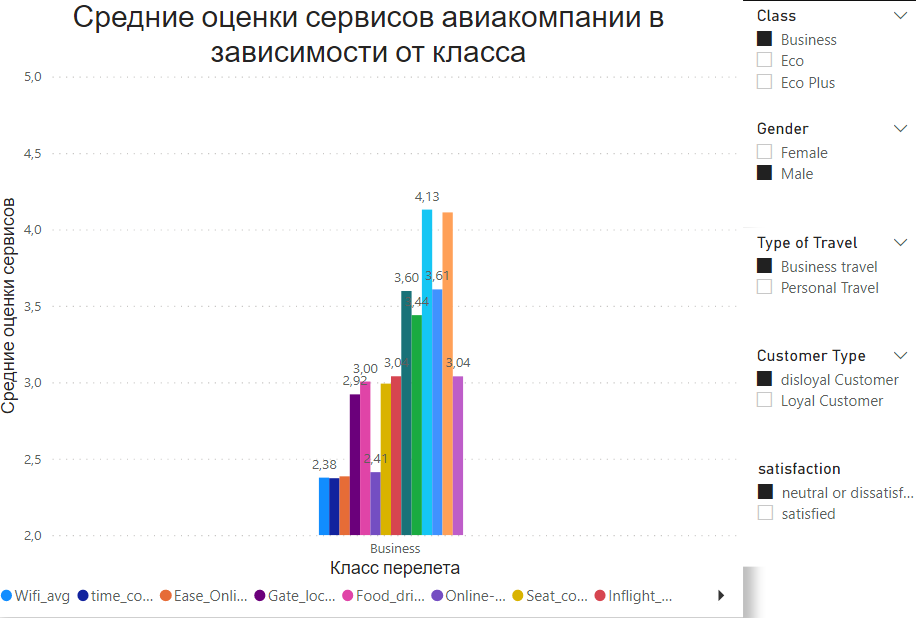


Рисунок 3.17 – Оценки пассажиров с использованием «Slicer» по категориям

Далее создадим ещё одну страницу для визуализации и из видов визуализации, представленных на рисунке 3.11 выберем тип – «Ribbon chart».

Настройки визуализации приведены на рисунке 3.18. Также аналогично были изменены заголовок, легенда и добавлены типы данных.

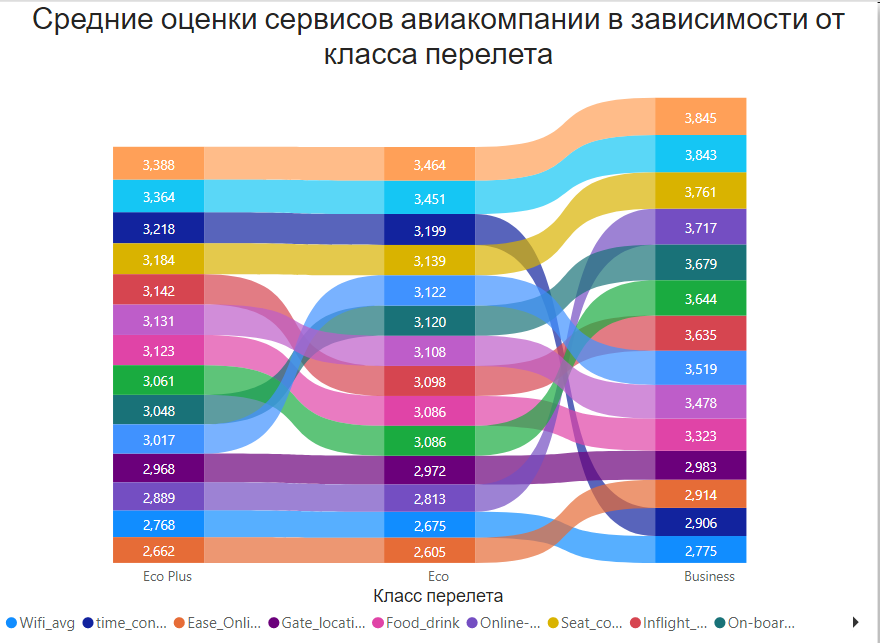


Рисунок 3.18 – Оценки пассажиров в зависимости от класса перелета

На данном рисунке можно наблюдать, в каком классе конкретный сервис имеет более высокие оценки среди пассажиров, а также как эти сервисы изменяются в зависимости от класса перелета. Также на диаграмму были добавлены аналогичные слайсеры.

Добавим еще одну страницу для визуализации, на которую поместим 4 диаграммы «Stacked column chart», где можно будет изучить количество удовлетворенности пассажиров по полу, классу перелета, типу путешествия и типу клиента. Результат представлен на рисунке 3.19.

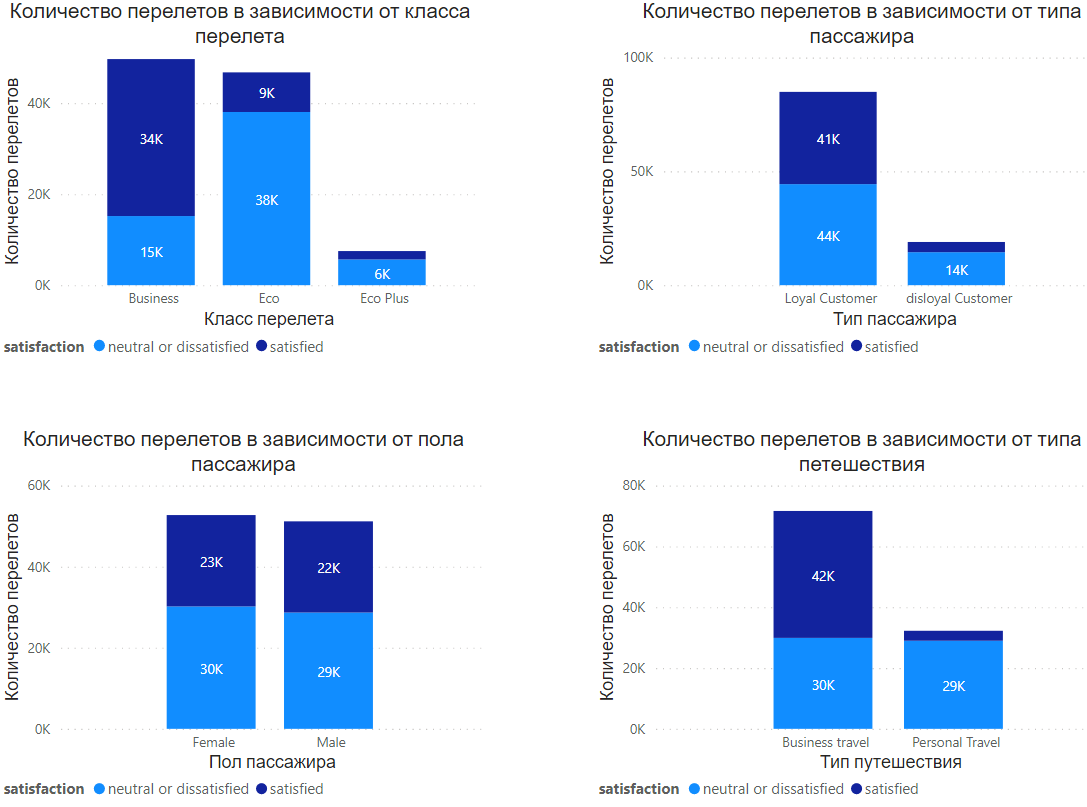


Рисунок 3.19 – Количество удовлетворенных и неудовлетворенных пассажиров по категориям

На рисунке 3.20 представлена визуализация доли удовлетворенных пассажиров в зависимости от класса перелета.

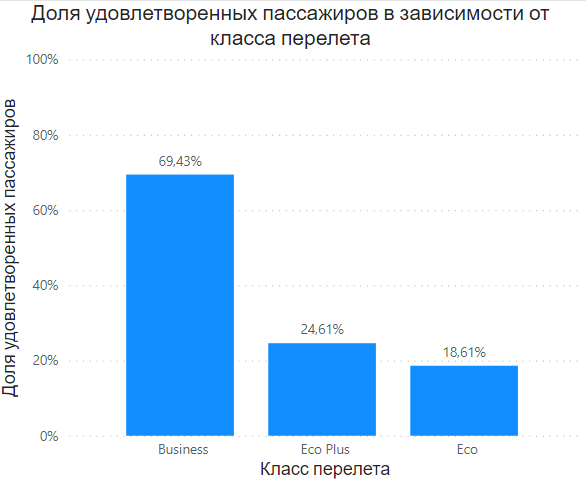


Рисунок 3.20 – Доля удовлетворенных пассажиров в зависимости от класса перелета

Исходя из визуализации, видно, что наиболее удовлетворенные пассажиры летали бизнес-классом, далее по удовлетворенности идет класс эко-плюс и затем эко.

Аналогичным образом создадим визуализацию по типу пассажиров, полу пассажиров и типу путешествия.

На рисунке 3.21 представлена визуализация доли удовлетворенных пассажиров в зависимости от типа пассажира.

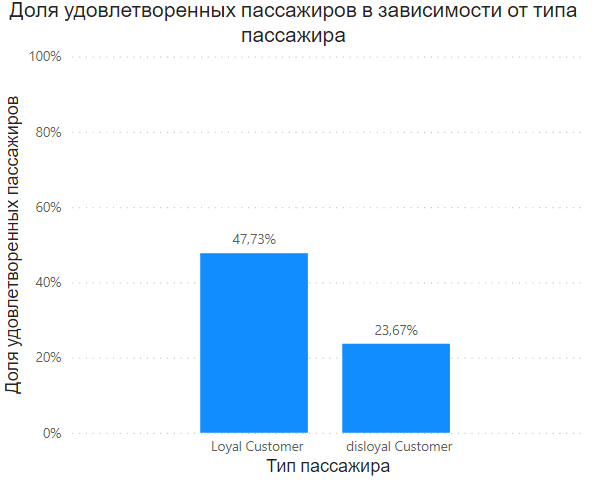


Рисунок 3.21 – Доля удовлетворенных пассажиров в зависимости от типа пассажира

Исходя из визуализации, видно, что наиболее удовлетворенные пассажиры, что очевидно, лояльные клиенты.

На рисунке 3.22 представлена визуализация доли удовлетворенных пассажиров в зависимости от пола пассажира.

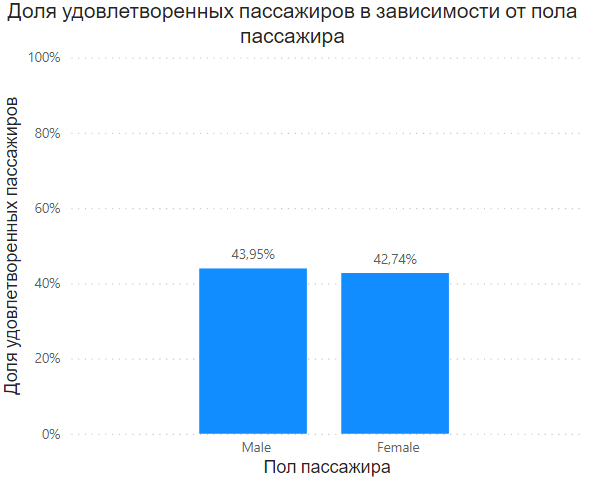


Рисунок 3.22 – Доля удовлетворенных пассажиров в зависимости от пола пассажира

Исходя из визуализации, видно, что нет разницы удовлетворенности пассажиров в зависимости от пола.

На рисунке 3.23 представлена визуализация доли удовлетворенных пассажиров в зависимости от типа путешествия.

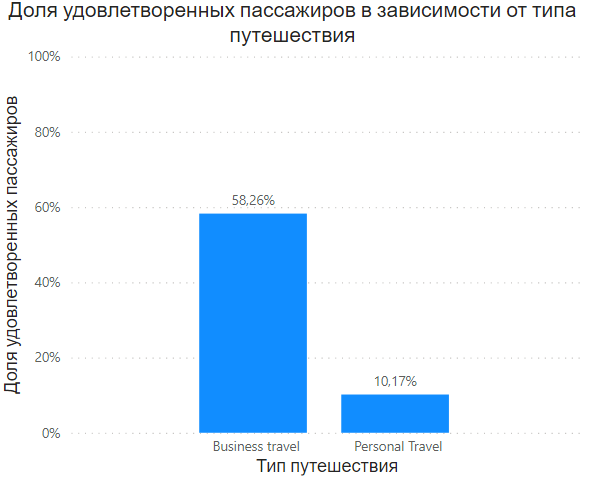


Рисунок 3.23 – Доля удовлетворенных пассажиров в зависимости от типа путешествия

Исходя из визуализации, видно, что удовлетворенных больше среди тех кто совершал перелет для рабочих целей.

Аналогично по удовлетворенности добавим диаграмму подсчета общего количества задержек рейсов при отправлении и прибытии в зависимости от класса перелета. Результат представлен на рисунке 3.24.

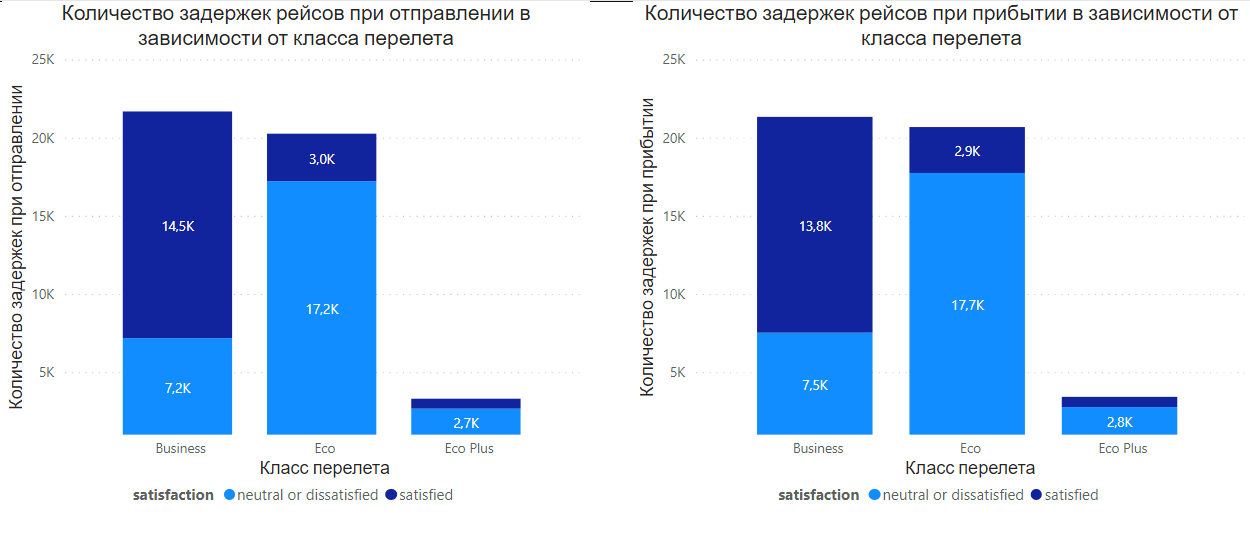


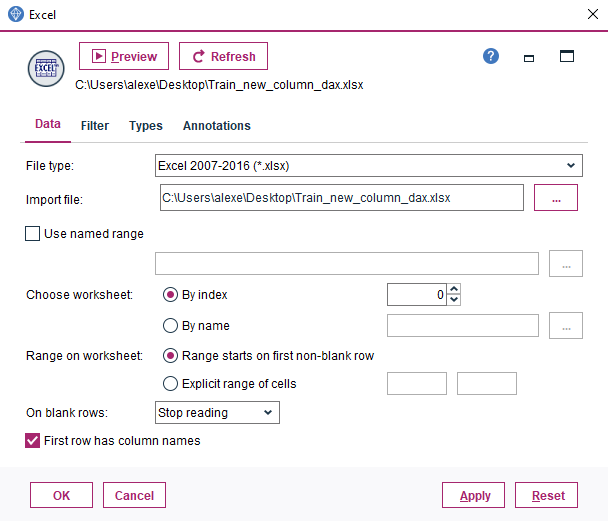
Рисунок 3.24 – Удовлетворенность пассажиров в зависимости от класса перелета по задержки рейса при отправлении и прибытии

# 4. АНАЛИЗ И ПОДГОТОВКА ДАННЫХ В IBM SPSS MODELER

Для импортирования файла с расширением «.xlsx» в IBM SPSS Modeler необходимо использовать ноду «Excel» из палитры «Sources» (рисунок 4.1).

  
Рисунок 4.1 – Нода «Excel»

В настройках Ноды «Excel» необходимо указать путь к файлу (рисунок 4.2).

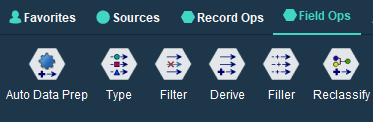
  
Рисунок 4.2 – Настройки для ноды «Excel»

Затем необходимо определить тип каждого поля в соответствии с таблицей 4.1.

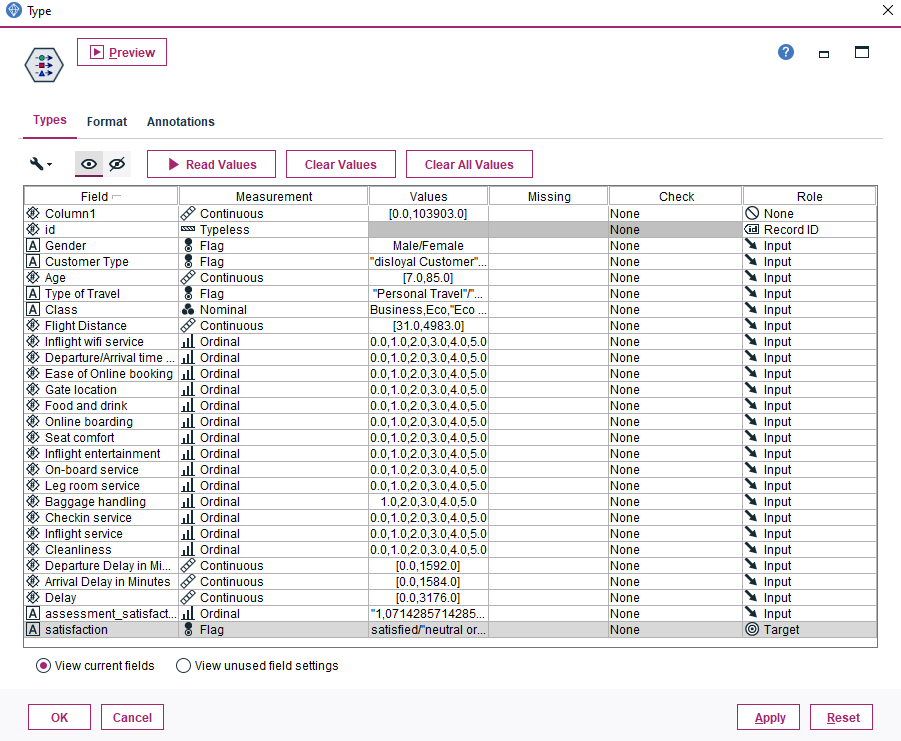
|  |  |  |
| --- | --- | --- |
| Атрибут | Описание атрибута | Тип поля |
| Column1 | Порядковый номер записи. | Continuous (Непрерывный).  *Будет удален из анализа.* |
| id | Идентификатор пассажира. | Typeless (Тип отсутствует) |
| Gender | Пол пассажира: женский, мужской. | Flag (Флаговый) |
| Customer Type | Тип пассажира: постоянный клиент, непостоянный клиент. | Flag (Флаговый) |
| Age | Возраст пассажира. | Continuous (Непрерывный) |
| Type of Travel | Цель полета пассажира (личная, деловая). | Flag (Флаговый) |
| Class | Класс в самолете (Бизнес, Эко, Эко Плюс). | Nominal (Номинативный) |
| Flight distance | Расстояние полета этого путешествия. | Continuous (Непрерывный) |
| Inflight wifi service | Уровень удовлетворенности услугой Wi-Fi на борту, где 0 - не приемлемо; 5 – замечательно.  (\*данная шкала будет использована в последующих атрибутах уровня удовлетворенности, кроме Baggage handling 1-5) | Ordinal (Упорядоченный) |
| Departure/Arrival time convenient | Уровень удовлетворенности удобным временем отправления / прибытия. | Ordinal (Упорядоченный) |
| Ease of Online booking | Уровень удовлетворенности легкостью онлайн-бронирования. | Ordinal (Упорядоченный) |
| Gate location | Уровень удовлетворенности местоположением выхода на посадку. | Ordinal (Упорядоченный) |
| Food and drink | Уровень удовлетворенности едой и напитками. | Ordinal (Упорядоченный) |
| Online boarding | Уровень удовлетворенности онлайн-посадкой. | Ordinal (Упорядоченный) |
| Seat comfort | Уровень удовлетворенности комфортом сиденья. | Ordinal (Упорядоченный) |
| Inflight entertainment | Уровень удовлетворенности развлечениями в полете. | Ordinal (Упорядоченный) |
| On-board service | Уровень удовлетворенности бортовым обслуживанием. | Ordinal (Упорядоченный) |
| Leg room service | Уровень удовлетворенности местом для ног. | Ordinal (Упорядоченный) |
| Baggage handling | Уровень удовлетворенности обработкой багажа. | Ordinal (Упорядоченный) |
| Check-in service | Уровень удовлетворенности службой регистрации. | Ordinal (Упорядоченный) |
| Inflight service | Уровень удовлетворенности обслуживанием в полете. | Ordinal (Упорядоченный) |
| Cleanliness | Уровень удовлетворенности чистотой в самолете. | Ordinal (Упорядоченный) |
| Departure Delay in Minutes | Задержка в минутах при отправлении. | Continuous (Непрерывный) |
| Arrival Delay in Minutes | Задержка в минутах при прибытии. | Continuous (Непрерывный) |
| Delay | Общая задержка рейса в минутах | Continuous (Непрерывный) |
| assessment\_satisfaction | Средняя оценка перелета по всем сервисам | Ordinal (Упорядоченный) |
| Satisfaction | Удовлетворенность авиакомпанией (удовлетворён, нейтрален или неудовлетворён). | Flag (Флаговый) |

Таблица 4.1 – Тип полей.

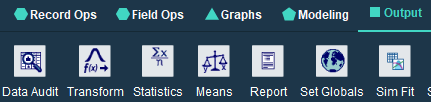
Для того чтобы назначить тип каждому полю, добавляем в поток ноду «Type» из палитры «Field Ops» (рисунок 4.3).

  
Рисунок 4.3 – Нода «Type»

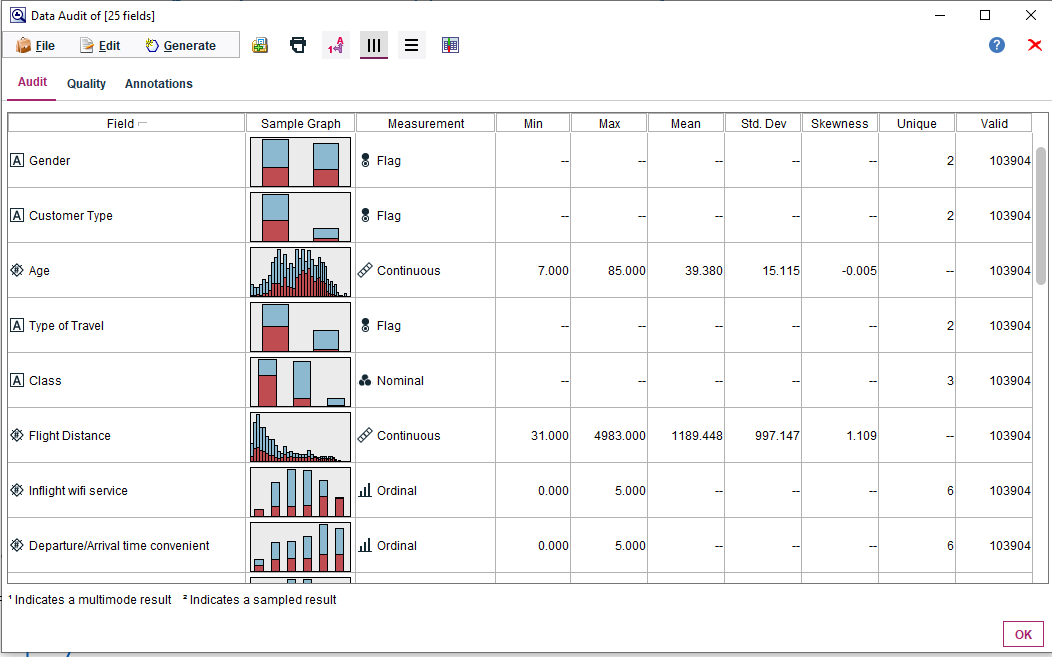
В настройках ноды «Type» указываем типы полей и нажимаем кнопку «Read values» (рисунок 4.4). Также указываем целевую переменную «satisfaction» и установить значение «None» для поля «Column1».

  
Рисунок 4.4 – Настройка типов полей

Далее перейдем к анализу качества импортируемых данных. Для этого воспользуемся нодой «Data audit» из палитры «Output» (рисунок 4.5).

  
Рисунок 4.5 – Нода «Data Audit»

Результат запуска ноды «Data audit» представлен на рисунке 4.6.

  
Рисунок 4.6 – Результат запуска ноды «Data Audit»

Далее на вкладке «Quality» можно проанализировать количество пропущенных значений, бланковых значений и пустых строк по каждому полю.

Таким образом, показатели качества по каждому полю представлены на рисунке 4.7.

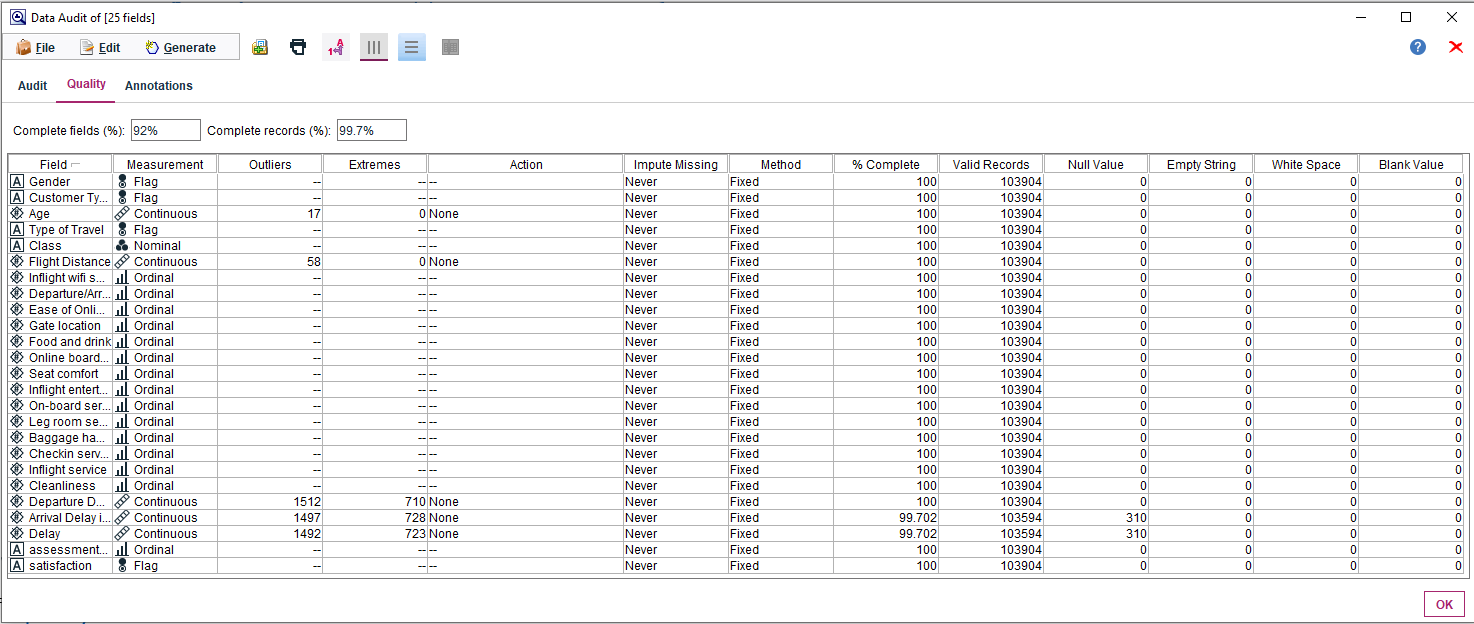


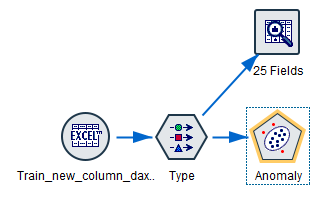
Рисунок 4.7 – Показатели качества для каждого поля

По результатам анализа качества исходных данных можно сделать следующий вывод: качество данных для построения моделей высокое. Однако стоит отметить, что будет проведено исследование данных на наличие аномалий, а также в некоторых моделях потребуется дополнительная обработка.

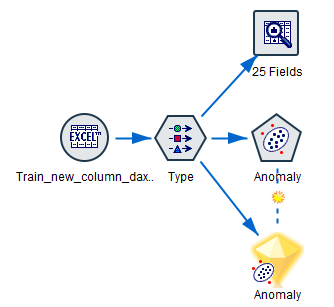
Также необходимо исследовать данные на наличие аномалий, для их будем использовать ноду «Anomaly» из палитры «Modeling» (рисунок 4.8).

  
Рисунок 4.8 – Нода «Anomaly»

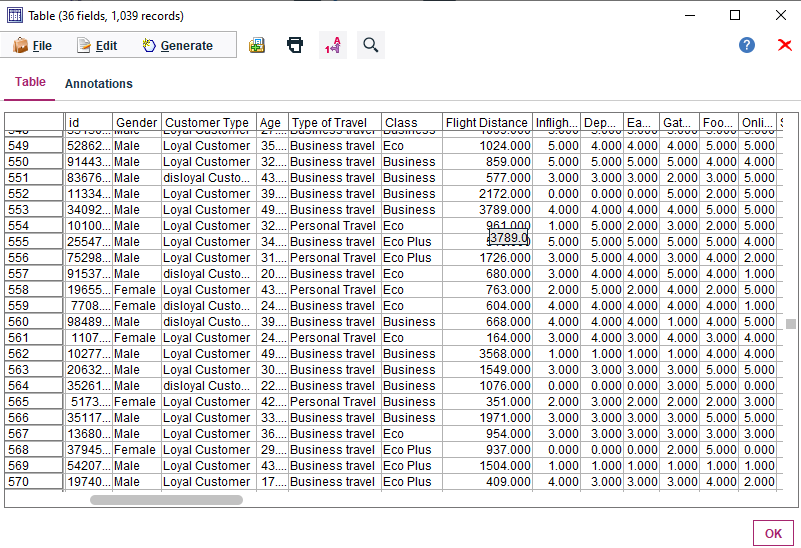
Схема потока представлена на рисунке 4.9.

  
Рисунок 4.9 – Схема потока

В результате запуска модели создается «Слепок модели» желтого цвета в виде самородка (в дальнейшем будет называть - «Nugget»), который представлен на рисунке 4.10.

  
Рисунок 4.10 – Схема потока после запуска ноды «Anomaly»

В настройках «Nugget» можно указать действия, которые мы хотим применить по отношению к аномальным данным. Отобразим эти данные с помощью ноды «Table» (рисунок 4.11).

  
Рисунок 4.11 – Список записей с аномальными значениями

Количество записей, содержащих аномальные значения равно 1039, удалим их с помощью опции «Discard records» в настройках слепка модели (рисунок 4.12).

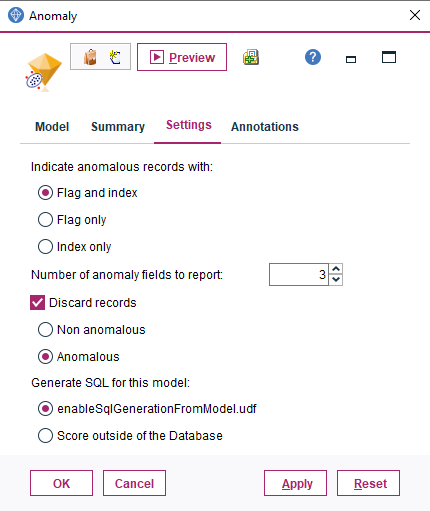


Рисунок 4.12 – Удаление аномальных данных

После проведенных действий датасет содержит 102 865 строк (рисунок 4.13).

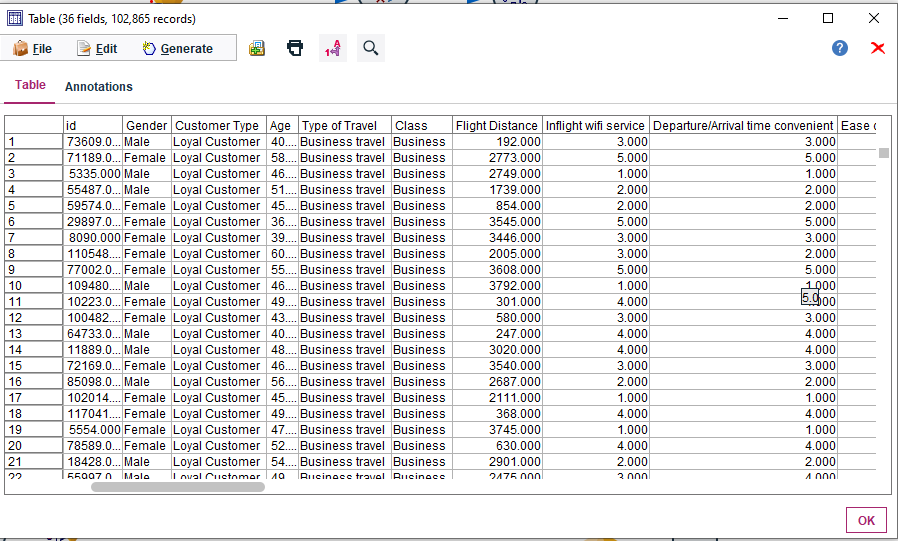


Рисунок 4.13– Датасет без аномальных данных

Чтобы исключить поля, созданные в ходе исследования данных на аномалии, добавим ноду «Filter» из палитры «Field Ops» (рисунок 4.14).

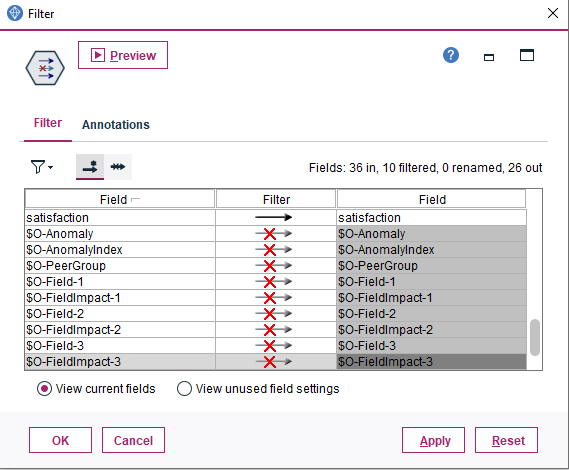


Рисунок 4.14 – Применение ноды «Filter»

Далее для сокращения размерности задачи проведем PCA-анализ, из палитры «Modeling». Схема потока после запуска PCA-анализа представлена на рисунке 4.15.

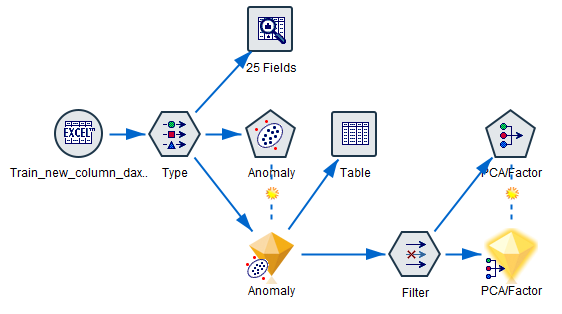


Рисунок 4.15 – Схема потока PCA-анализа

Результаты PCA-анализа представлены на рисунке 4.16. В столбце «Extraction» показано, какая доля дисперсии приходится на извлеченные компоненты. В качестве примера, так как значения в целом высокие уберем из анализа поля, для которых «Extraction» имеет меньше 0.4.

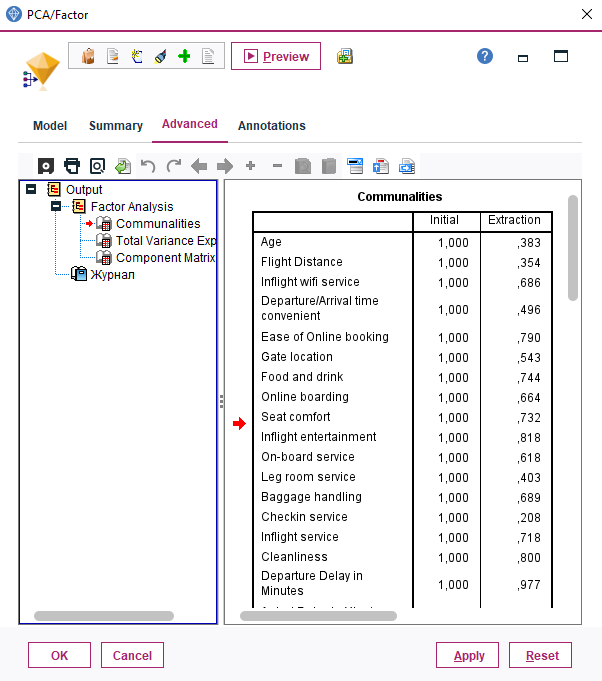


Рисунок 4.16 – Результат применения PCA-анализа

Удалим из анализа следующие поля (рисунок 4.17):

* Пол;
* Дистанцию перелета;
* Класс пассажира;
* Тип путешествия;
* Тип пассажира;
* Возраст;
* Уровень удовлетворенности местом для ног;
* Уровень удовлетворенности местом службой регистрации.

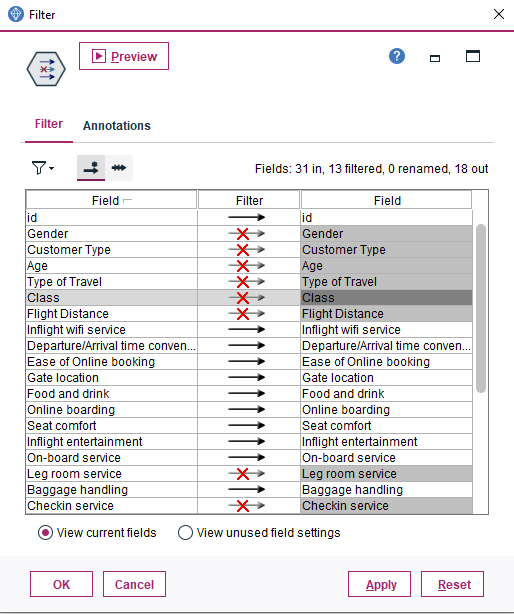


Рисунок 4.17 – Удаление полей после PCA-анализа

# 5. МОДЕЛИРОВАНИЕ И ОЦЕНКА КАЧЕСТВА МОДЕЛИ

На этапе моделирования нужно определить возможные модели, которые наиболее подходят для решения поставленной задачи и учитывают особенности имеющихся данных.

С целью определения подходящих моделей был проведен сравнительный анализ, результаты которого представлены в таблице 5.1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Показатель | QUEST | C5.0 | C&R | CHAID | RandomForest | Logistic | Neural Net |
| Обработка пропущенных значений | Использование суррогатных полей | Фракционализация | Использование суррогатных полей | Создание отдельной категории | Да | Нет | Нет |
| Бэггинг | Да | Нет | Да | Да | - | - | - |
| Возможность стрижки | Да | Да | Да | Нет | Да | - | - |
| Бинарное дерево | Да | Нет | Да | Нет | - | - | - |
| Подходит для задачи бинарной классификации | Да | Да | Да | Да | Да | Нет | Да |
| Интерпретируемость модели | Да | Да | Да | Да | Нет | Да | Нет |

Таблица 5.1 – Сравнительная характеристика моделей

Целевая переменная «satisfaction» представлена флаговым типом, а предикторы являются как категорными так и не прерывными будут использованы следующие модели:

* CHAID Tree;
* C&R Tree;
* C5.0 Tree;
* Quest Tree;
* RandomForest.

В качестве метрик для оценки качества модели бинарной классификации были выбраны следующие метрики:

* Accuracy;
* AUC (Area Under Curve).

AUC – площадь под ROC кривой, показывает вероятность того, что случайно выбранный экземпляр негативного класса будет иметь меньше вероятность быть распознанным как позитивный класс, чем случайно выбранный позитивный класс (чем ближе к 1, тем лучше).

Начнем с модели «CHAID» для этого из палитры «Modeling» добавим в поток ноду «CHAID» и запустим её. Затем к слепку модели (Nugget) добавим ноду «Analysis», предварительно установив вывод информации о качестве модели (рисунок 5.1).

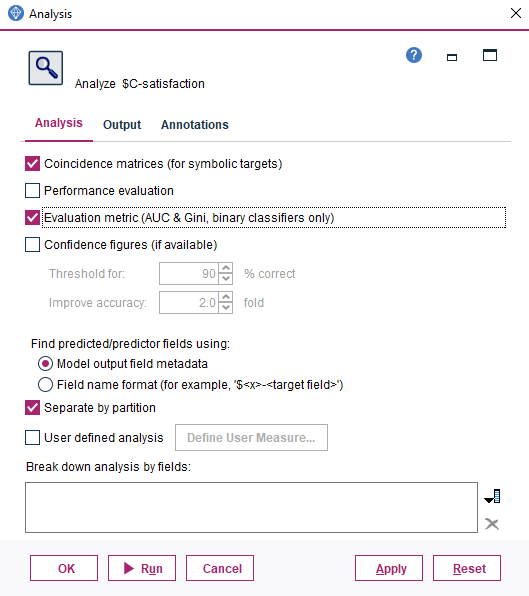


Рисунок 5.1 – Качество модели, нода «Analysis»

Запустив ноду «Analysis», мы получим данные о точности и качестве модели «CHAID» (рисунок 5.2).

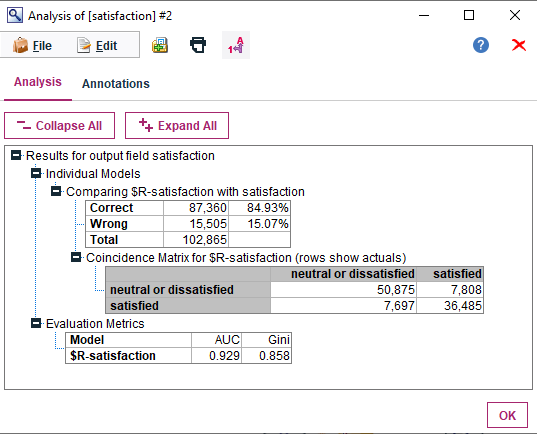


Рисунок 5.2 – Результат запуска ноды «Analysis» модели «CHAID»

Как видно из результатов точность модели оказалась довольно высокой – 84.93%, AUC – 0.929.

На рисунке 5.3 представлена важность предикторов для модели «CHAID».

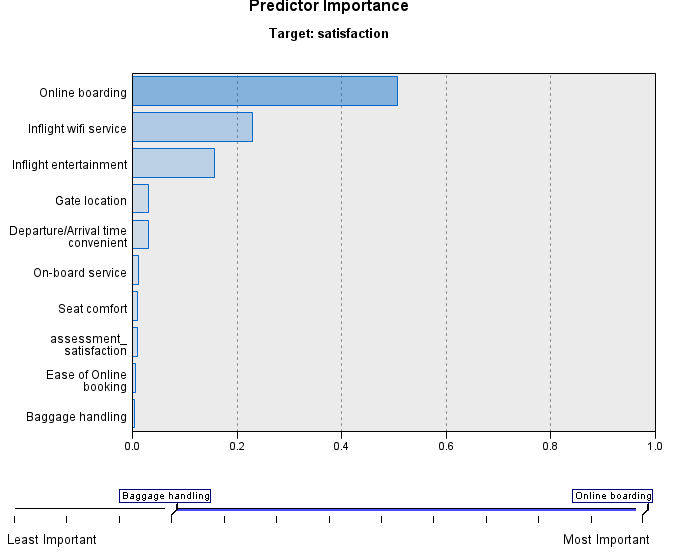


Рисунок 5.3 – Важность предикторов для модели «CHAID»

Наиболее важными предикторами являются:

* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности Wi-Fi;
* Уровень удовлетворенности развлечениями на борту;

Перейдем к следующему виду модели, для этого аналогичным образом из палитры «Modeling» добавим в поток ноду «C&R Tree» - запустим её, после аналогично добавим ноду «Analysis» и снова запустим (рисунок 5.4).

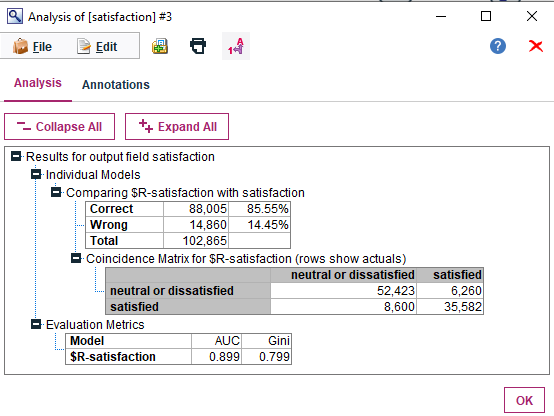


Рисунок 5.4 – Результат запуска ноды «Analysis» модели «C&R Tree»

Как видно из результатов точность модели - 85.55%, AUC – 0.899.

На рисунке 5.5 представлена важность предикторов для модели «C&R».

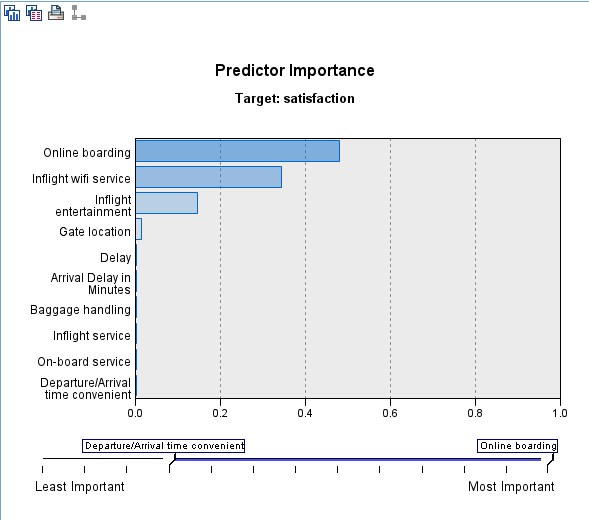


Рисунок 5.5 – Важность предикторов для модели «C&R»

Наиболее важными предикторами являются:

* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности Wi-Fi;
* Уровень удовлетворенности развлечениями на борту;

Перейдем к следующему виду модели, для этого аналогичным образом из палитры «Modeling» добавим в поток ноду «C5.0» - запустим её, после аналогично добавим ноду «Analysis» и снова запустим (рисунок 5.6).

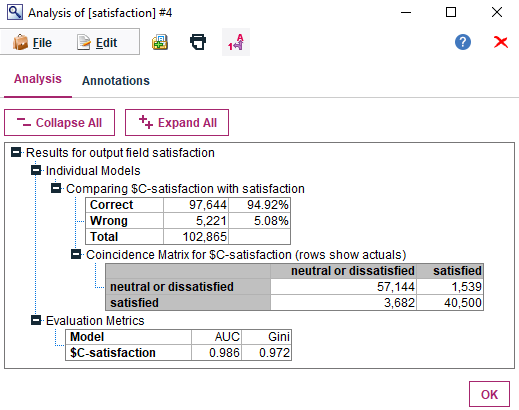


Рисунок 5.6 – Результат запуска ноды «Analysis» модели «C5.0»

Как видно из результатов точность модели - 94.92%, AUC – 0.986.

Наиболее важными предикторами являются (рисунок 5.7):

* Уровень удовлетворенности Wi-Fi;
* Уровень удовлетворенности выходом на посадку;
* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности временем отправления и прибытия;
* Уровень удовлетворенности развлечениями на борту;

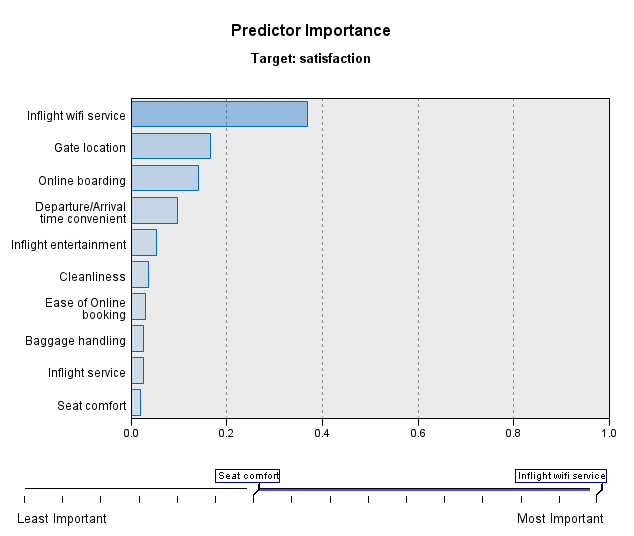


Рисунок 5.7 – Важность предикторов для модели «C5.0»

Перейдем к следующему виду модели, для этого аналогичным образом из палитры «Modeling» добавим в поток ноду «Quest» - запустим её, после аналогично добавим ноду «Analysis» и снова запустим (рисунок 5.8).

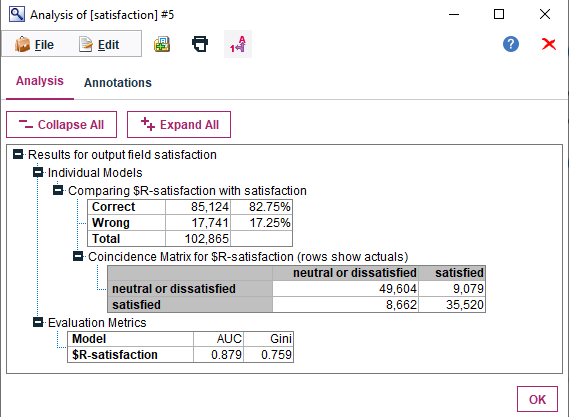


Рисунок 5.8 – Результат запуска ноды «Analysis» модели «Quest»

Как видно из результатов точность модели - 82.75%, AUC – 0.879.

Наиболее важными предикторами являются (рисунок 5.9):

* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности развлечениями на борту;
* Уровень удовлетворенности бортовым обслуживанием.

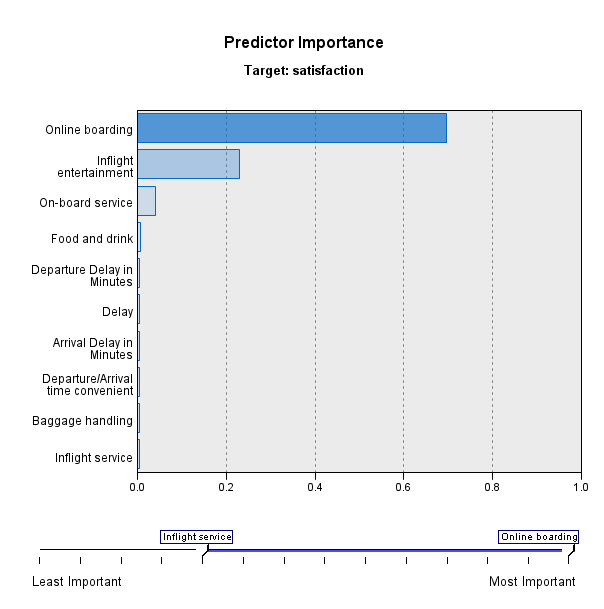


Рисунок 5.9 – Важность предикторов для модели «Quest»

Перейдем к другому виду модели для этого аналогичным образом из палитры «Modeling» добавим в поток ноду «Random Forest» - запустим её, после аналогично добавим ноду «Analysis» и снова запустим (рисунок 5.10).

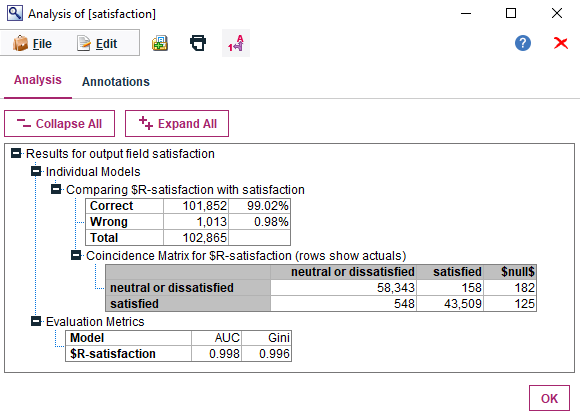


Рисунок 5.10 – Результат запуска ноды «Analysis» модели «Random Forest»

Как видно из результатов точность модели оказалась высокой, выше, чем у предыдущих моделей, точность – 99,02%, AUC – 0,998.

Возможно, ввиду невозможности интерпретации модели «Random Forest» предпочтение будет отдано модели «C5.0», однако все модели будут протестированы на тестовой выборке.

В таблице 5.2 в порядке убывания точности представлены используемые модели.

|  |  |
| --- | --- |
| Название модели | Точность модели |
| Random Forest | 99.02% |
| C5.0 Tree | 94.92% |
| C&R Tree | 85.55% |
| CHAID Tree | 84.93% |
| Quest Tree | 82.75% |

Таблице 5.2 – точность используемых моделей.

Общая схема потока представлена на рисунке 5.11

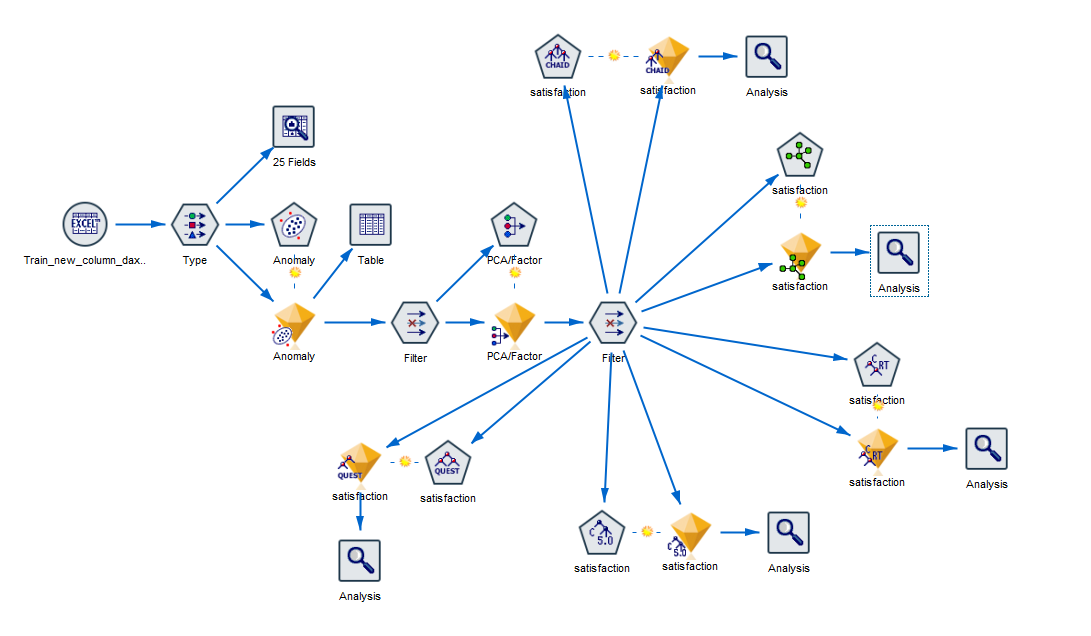


Рисунок 5.11 – Итоговая схема потока в IBM SPSS Modeler

# 6. ПРОВЕРКА МОДЕЛЕЙ НА ТЕСТОВЫХ ДАННЫХ

Приступим к проверке моделей на тестовых данных, для этого с помощью ноды «Excel» загрузим тестовые данные. Далее с помощью ноды «Type» определим типы полей и уберем из анализа целевую переменную (рисунок 6.1).

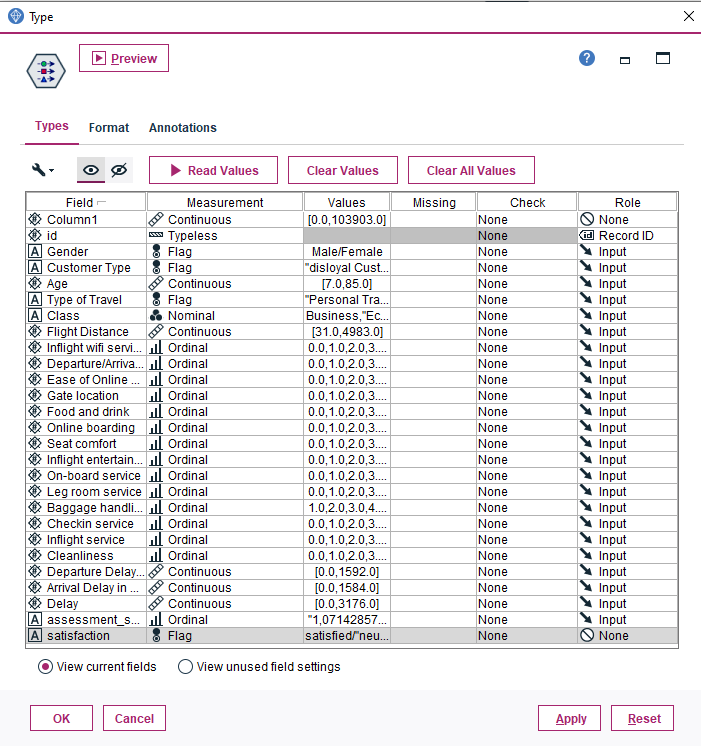


Рисунок 6.1 – Определение типов полей для тестовых данных

Повторим аналогичные действия по запуску анализа со всеми моделями, получим соответствующие результаты точности и качества моделей на тестовых данных.

Результат анализа точности и качества модели «Quest» представлен на рисунке 6.2.

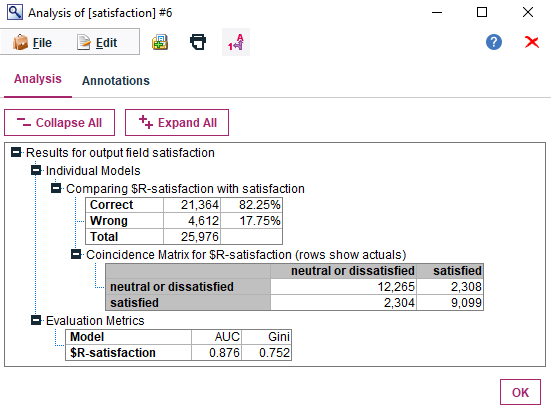


Рисунок 6.2 – Точность и качество модели «Quest»

Результат анализа точности и качества модели «C5.0» представлен на рисунке 6.3.

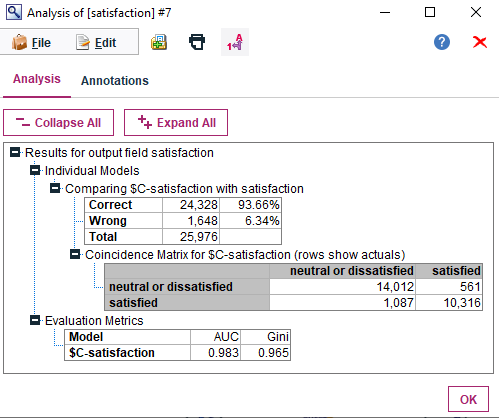


Рисунок 6.3 – Точность и качество модели «C5.0»

Результат анализа точности и качества модели «C&R» представлен на рисунке 6.4.

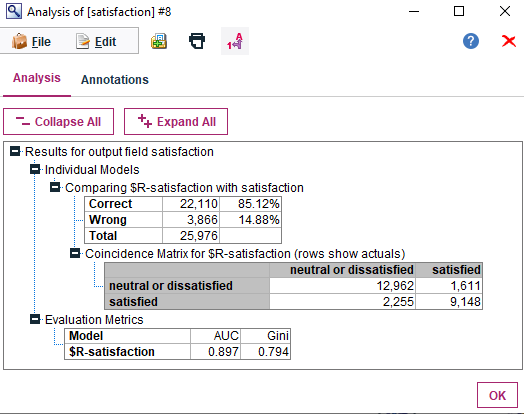


Рисунок 6.4 – Точность и качество модели «C&R»

Результат анализа точности и качества модели «CHAID» представлен на рисунке 6.5.

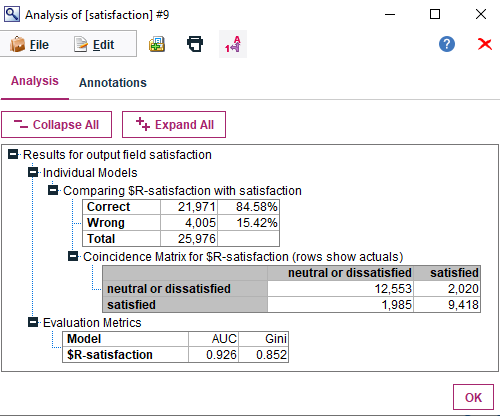


Рисунок 6.5 – Точность и качество модели «CHAID»

Результат анализа точности и качества модели «Random Forest» представлен на рисунке 6.6.

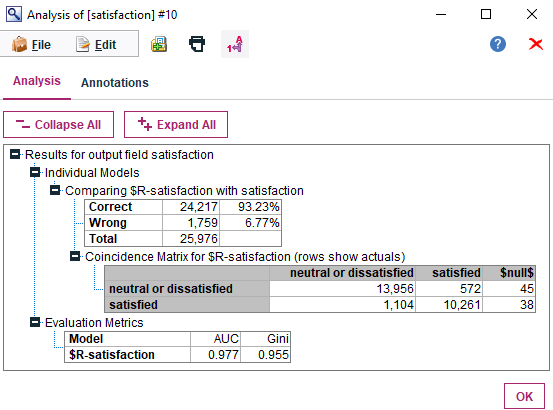


Рисунок 6.6 – Точность и качество модели «Random Forest»

Наибольшей точностью обладает модель С5.0, таким образом, в качестве наилучшей модели было выбрано «C5.0 Tree», метрики качества для которой следующие:

* Accuracy (Training) = 94.92%;
* Accuracy (Testing) = 93.66%;
* AUC (Training) = 0.986;
* AUC (Testing) = 0.983.

Важность предикторов модели «C5.0 Tree» представлена на рисунке 6.7.

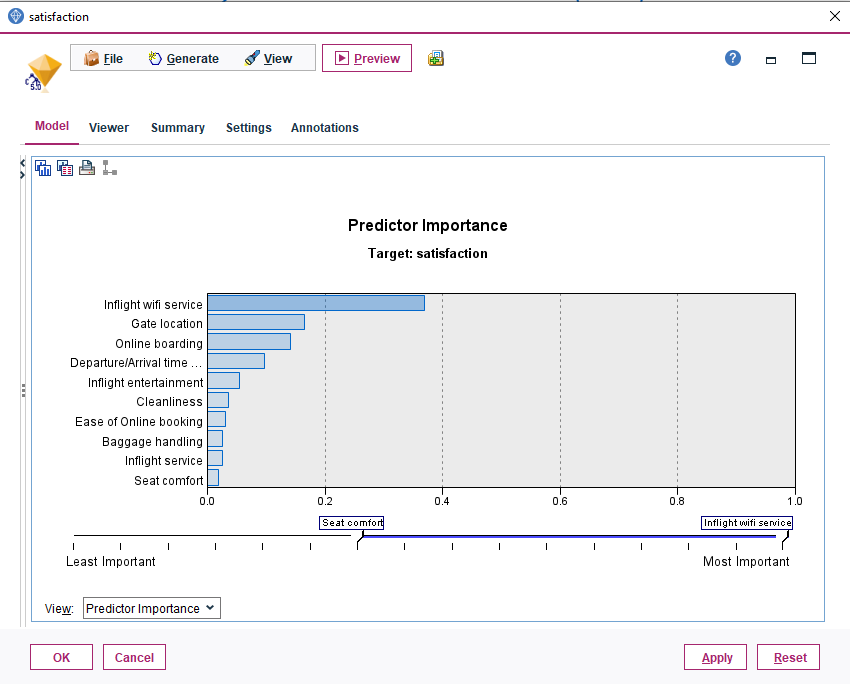


Рисунок 6.7 – Степень важности предикторов модели «C5.0 Tree»

Наиболее важными предикторами модели «C5.0 Tree» являются:

* Уровень удовлетворенности Wi-Fi;
* Уровень удовлетворенности выходом на посадку;
* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности временем отправления и прибытия;
* Уровень удовлетворенности развлечениями на борту;
* Уровень удовлетворенности чистотой в самолете.

# 7. ВЫВОДЫ ПО РАБОТЕ

В результате анализа верхнего уровня выявлено, что среди представленных классов перелета средние оценки по предоставлению, например, вай-фай услуг соизмеримы между классами перелета.

Пассажиры всех классов неудовлетворенны данным сервисом. Также стоит отметить низкие средние оценки легкости онлайн-бронирования, возможно, авиакомпании следует уделить этому внимание и выяснить причину таких оценок.

Возможно, принять меры по введению инструкций по бронированию, устранения каких либо ошибок, недочетов, четкой структуризации этапов бронирования билетов.

Наиболее высокие оценки при перелете получили следующие сервисы авиакомпании:

* Бортовое обслуживание в самолете;
* Комфортность сидений;
* Обработка багажа.

Были разработаны модели прогнозирования удовлетворенности пассажиров перелетом. По результатам точности выбрана самая лучшая по точности и качеству модели – C5.0 Tree (точность 93.66%, AUC = 0.983).

Итоговая схема потока представлена на рисунке 7.1.

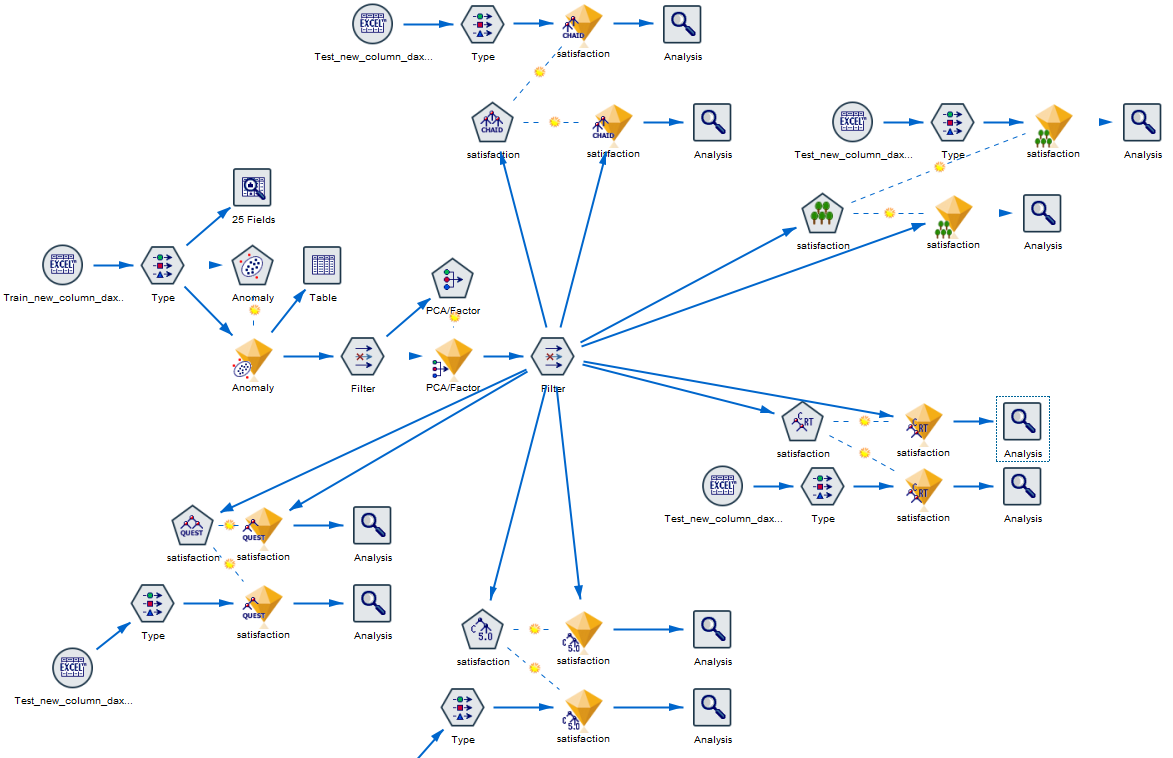


Рисунок 7.1 – Общая схема потоков моделей

Во время глубинного анализа были выяснены наиболее важные предикторы при выставлении оценок пассажирами, а именно:

* Уровень удовлетворенности Wi-Fi сервисом;
* Уровень удовлетворенности онлайн-посадкой;
* Уровень удовлетворенности выходом на посадку;
* Уровень удовлетворенности бортовым обслуживанием;
* Уровень удовлетворенности развлечениями во время полета;
* Уровень удовлетворенности чистотой в самолете.

В ходе работы выполнены следующие задачи:

* Проведена работа по обогащению данных (удалены некоторые столбцы, добавлены столбцы в качестве вычислимых полей, добавлены таблицы, созданные на основе первичной).
* Проведен верхнеуровневый анализ в MS Power BI;
* Построены графики визуализации данных;
* Проведен глубинный анализ в IBM SPSS Modeler;
* Использованы и описаны 5 видов моделей предсказания удовлетворенности пассажиром авиаперелетом;
* Проведена проверка моделей на тестовых данных;
* Выбрана лучшая модель.