Федеральное государственное автономное образовательное учреждение

высшего образования

«Национальный исследовательский университет «Высшая школа экономики»

Высшая школа бизнеса

Образовательная программа «Бизнес-информатика»

О Т Ч Е Т

по проектной работе

«Моделирование своевременности доставки товаров транспортно-логистической компании с использованием предиктивной аналитики на базе IBM SPSS Modeler»

Выполнил студент гр. МБИ-202

Андросов Алексей Константинович

**Руководитель проекта:**

Брускин Сергей Наумович,

доцент департамента бизнес-

информатики, к.э.н.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(оценка)*

*\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(Дата) (подпись)*

Москва 2021

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc89931448)

[1. ПОСТАНОВКА БИЗНЕС-ЗАДАЧИ 6](#_Toc89931449)

[1.1. Описание бизнес-кейса 6](#_Toc89931450)

[1.2. Общее описание данных, доступных для исследования 7](#_Toc89931451)

[1.3. Предварительный анализ с использованием MS Power BI 9](#_Toc89931452)

[1.4. Основные гипотезы исследования 23](#_Toc89931453)

[1.5. Требования к результатам 26](#_Toc89931454)

[1.6. Методология, роли и этапы проведения исследования 26](#_Toc89931455)

[1.7. Выводы по результатам Этапа 1 28](#_Toc89931456)

[2. АНАЛИЗ И ПОДГОТОВКА ДАННЫХ 30](#_Toc89931457)

[2.1. Теоретический обзор приложений предиктивной аналитики 30](#_Toc89931458)

[2.2. Планируемые методы исследования 32](#_Toc89931459)

[2.3. Импортирование и первичный анализ данных 33](#_Toc89931460)

[2.4. Оценка качества данных 35](#_Toc89931461)

[2.5. Сценарии сокращения размерности задачи 39](#_Toc89931462)

[2.6. Выводы по результатам Этапа 2. 42](#_Toc89931463)

[3. МОДЕЛИРОВАНИЕ И ОЦЕНКА КАЧЕСТВА МОДЕЛИ 43](#_Toc89931464)

[3.1. Настройка модели и описание применяемых методов прогнозирования 43](#_Toc89931465)

[3.2. Описание сценариев и оценка качества модели 45](#_Toc89931466)

[3.3. Интерпретация полученных результатов 53](#_Toc89931467)

[3.4. Выводы по результатам Этапа 3 54](#_Toc89931468)

[4. ОЦЕНКА РЕЗУЛЬТАТОВ ПРОЕКТА 56](#_Toc89931469)

[4.1. Описание результатов проекта 56](#_Toc89931470)

[4.2. Описание сформированных компетенций 57](#_Toc89931471)

[4.3. Оценка перспектив развития полученных решений 58](#_Toc89931472)

[ЗАКЛЮЧЕНИЕ 59](#_Toc89931473)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 61](#_Toc89931474)

# ВВЕДЕНИЕ

Глоссарий:

* Проект – структурно оформленное и логически завершенное прикладное исследование по выбранной тематике, направленное на достижение цели в соответствии с функциональными и специальными требованиями, проектными критериями и графиком реализации;
* Постановка задачи – краткое описание задачи исследования в форме бизнес-кейса, включающее описание объекта исследования, наборов данных, используемых для анализа, планируемых методов и инструментов исследования, а также ожидаемых результатов;
* Методология реализации проекта – подходы и методики проведения исследования в рамках Проекта, логически объединенные в единую систему для поддержки всех этапов и оформления результатов проектной деятельности;
* IBM CRISP-DM – методология исследования, принятая в качестве стандарта для реализации проекта по теме «Прогнозное моделирование показателей результативности в различных сферах бизнеса с использованием предиктивной аналитики»;
* Предиктивная аналитика (predictive analytics) — класс методов интеллектуального анализа данных, использующих статистические и кибернетические инструменты моделирования, ориентированный на прогнозирование будущего состояния исследуемых объектов и направленный на улучшение качества принимаемых решений;
* Показатели результативности (performance indicators) – значимые показатели поведения объекта исследования, описанные в рамках Постановки задачи в качестве целевых переменных для прогнозного моделирования.

Целью работы является разработка и внедрение модели прогнозирования своевременности доставки товаров клиентам крупной транспортно-логистической компании.

Объектом исследования является компания в сфере грузоперевозок – «DEL».

Предметом исследования являются факторы, влияющие на своевременность доставки товаров компании «DEL».

Тип проекта: исследовательский.

Руководитель проекта: Брускин Сергей Наумович, к.э.н., доцент, [sbruskin@hse.ru](mailto:sbruskin@hse.ru).

Процесс выполнения проекта будет разбит на следующие этапы:

* Постановка бизнес-задачи;
* Анализ и подготовка данных;
* Моделирование и оценка качества моделирования;
* Защита проекта.

В ходе исследования для реализации вышеуказанных этапов перед нами стоят следующие задачи:

* Описать бизнес-задачу проекта;
* Провести верхнеуровневый, визуальный анализ с использованием MS Power BI;
* Сформулировать гипотезы исследования;
* Провести анализ, очистку и оценку данных;
* Разработать модель для предсказания своевременности доставки товаров компьютерной периферии клиентам;
* Описать используемые методы прогнозирования;
* Описать сценарии;
* Оценить качество модели;
* Проинтерпретировать полученные результаты исследования;
* Оценить перспективы развития полученных решений;
* Описать результаты исследования в форме отчета.

# 1. ПОСТАНОВКА БИЗНЕС-ЗАДАЧИ

## 1.1. Описание бизнес-кейса

«DEL» является австралийской компанией, которая занимается грузоперевозками как внутри страны, так и за ее пределами. В распоряжение компании находится 5 складов:

* склад «A» - Новый Южный Уэльс;
* склад «B» - Квинсленд;
* склад «C» - Западная Австралия;
* склад «D» - Виктория;
* склад «F» - Тасмания.

Компания использует следующие способы доставки товаров:

* авиаперевозки;
* морские перевозки;
* сухопутные перевозки.

В качестве объекта перевозок выступают товары компьютерной периферии.

Получение качественного продукта в требуемые сроки – ответственная и сложная задача, особенно в настоящих условиях пандемии короновируса, клиенты стали чаще заказывать товары и компаниям следует сделать все возможное, чтобы не только сохранить, но и приумножить количество своих клиентов.

Поэтому, когда региональная компания «DEL» по доставке продукции стала получать множество жалоб от клиентов из-за несвоевременности доставки, компания всерьез обеспокоилась этим и решила обратиться к ключевым сведениям из своей базы данных клиентов для выявления причин сложившейся ситуации.

Целью проекта является использование разработанной модели предиктивной аналитики для выявления факторов, наиболее сильно влияющих на своевременность доставки. Анализ данных факторов позволит компании «DEL» принять меры для снижения количества несвоевременных доставок.

## 1.2. Общее описание данных, доступных для исследования

В качестве данных был использован датасет «E-Commerce Shipping Data», взятый с сайта kaggle.com [5].

Фрагмент данных представлен на рисунке 1.1.

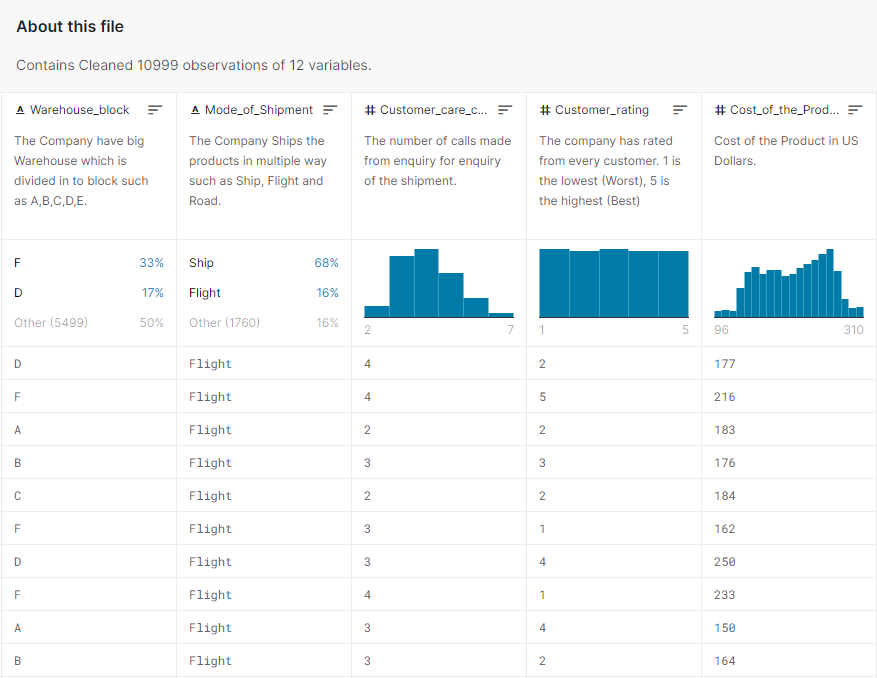


Рисунок 1.1. Фрагмент данных «E-Commerce Shipping Data»

Данные представлены тренировочным набором, данные для тестовой выборки будут сформированы случайным образом из тренировочного набора (~30%). В тренировочном наборе 11 тысяч записей и 12 полей. Целевая переменная – «**Reached.on.Time\_Y.N**», где 1 означает, что продукт не был доставлен вовремя, а 0, что он был доставлен вовремя. Период сбора данных составляет с марта по август 2021 года (6 месяцев). Список полей данных представлен в таблице 1.1.

Таблица 1.1 – Список полей данных.

|  |  |  |
| --- | --- | --- |
| **Название атрибута данных** | **Пояснение к названию** | **Перевод** |
| ID Number of Customers. | ID | Идентификатор поставки. |
| Warehouse\_block | The Company have big Warehouse which is divided in to block such as A,B,C,D,F. | Блоки склада: A,B,C,D,F. |
| Mode\_of\_Shipment | The Company Ships the products in multiple way such as Ship, Flight and Road. | Вид перевозки (самолет, поезд, автобус). |
| Customer\_care\_calls | The number of calls made from enquiry for enquiry of the shipment. | Кол-во звонков от клиента. |
| Customer\_rating | The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best) | Рейтинг клиента (1 - самый низкий (худший), 5 - самый высокий (лучший)). |
| Cost\_of\_the\_Product | Cost of the Product in US Dollars. | Стоимость продукта. |
| Prior\_purchases | The Number of Prior Purchase. | Количество предшествующих покупок. |
| Product\_importance | The company has categorized the product in the various parameter such as low, medium, high. | Классификация продукта по важности - низкий, средний, высокий. |
| Gender | Male and Female. | Пол клиента (Мужской/Женский). |
| Discount\_offered | Discount offered on that specific product. | Скидка на конкретный продукт. |
| Weight\_in\_gms | It is the weight in grams. | Вес продукта в граммах. |
| Reached.on.Time\_Y.N | It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time. | **Целевая переменная**, где 1 означает, что заказ НЕ был доставлен вовремя, а 0, что он был доставлен вовремя. |

## 1.3. Предварительный анализ с использованием MS Power BI

Для проведения предварительного визуального анализа будет использован Microsoft Power BI. Запускаем Microsoft Power BI, далее выбираем «Get data» 🡪 «Text/CSV» и через обзор находим необходимую папку с данными и подгружаем данные в формате .csv.

Предварительный просмотр загруженных данных представлен на рисунке 1.2.

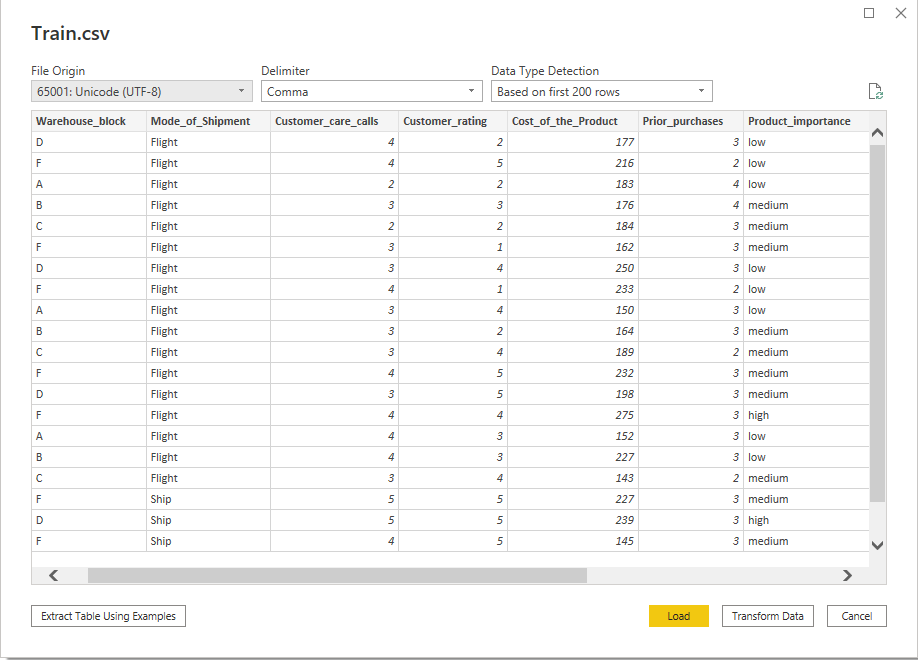


Рисунок 1.2 – Предварительный просмотр данных в Power BI

На данном шаге оставляем настройки по-умолчанию, затем переходим в пункт «Transform Data» и проводим следующие действия – меняем тип данных у столбца «id» с «Whole number» на «Text» (чтобы Power BI не стал суммировать идентификаторы клиентов).

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

Посчитаем количество отправленных товаров с каждого склада. Для этого сгруппируем данные по доставке с каждого склада компании «DEL», воспользовавшись операцией «Group By», настройки команды приведены на рисунке 1.3.

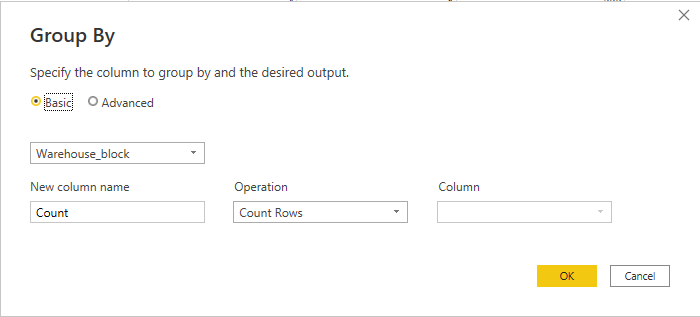


Рисунок 1.3 – Настройки операции «Group By»

Далее выполним сложную группировку по своевременности доставки с каждого склада. Для этого в первое поле группировки установим группировку по складу, а во второе – своевременность доставки. Настройки приведены на рисунке 1.4.

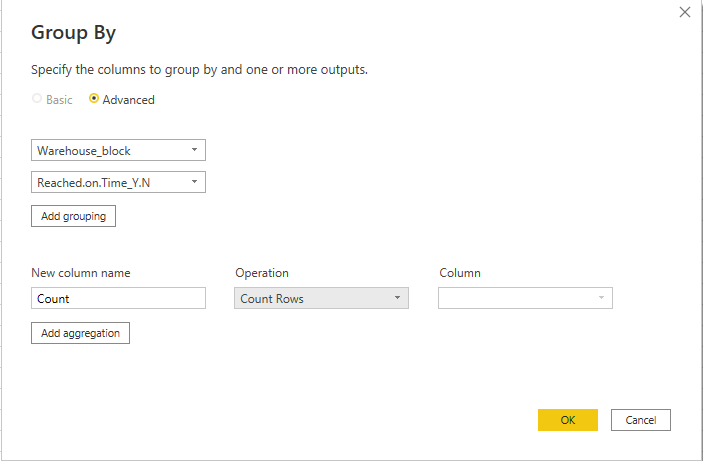


Рисунок 1.4 – Настройки сложной группировки «Group By»

В случае если будет необходимо выполнить операцию «Group By» снова, можно будет вернуться к трансформированию данных. Далее загружаем модифицированные данные в Power BI, нажав «Close & Apply».

Следующим шагом будет создание мер через операцию «New measure», создадим меру со средней стоимости покупки, в дальнейшем можно будет рассматривать среднюю стоимость как для тех покупателей, которым товар был доставлен несвоевременно, так и для тех, кому доставили товар вовремя. Формула расчета средней стоимости товаров приведена на рисунке 1.5.



Рисунок 1.5 – Формула расчета средней стоимости продуктов

Аналогичным образом, создадим следующие меры:

* Среднее количество звонков по отгрузке товаров;
* Среднее скидка по товарам;
* Среднее число предыдущих покупок;
* Средний рейтинг клиентов;
* Средний вес товаров.

Перечень, созданных мер представлен на рисунке 1.6.

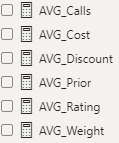


Рисунок 1.6 – Созданные меры

Приступим к выполнению первичного анализа и некоторой визуализации данных. Рассмотрим визуализацию стоимости продуктов отправленных с каждого из складов компании «DEL». Необходимо перейти в раздел «Report» и выбрать следующую визуализацию – «Treemap» (третий ряд, шестой элемент), виды визуализации представлены на рисунке 1.7.

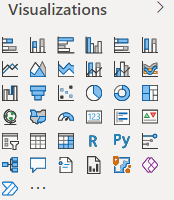


Рисунок 1.7 – Виды визуализации

Далее проведем настройку визуализации для этого необходимо нажать на саму визуализацию и перейти во вкладку «Fields». Настройки визуализации представлены на рисунке 1.8.

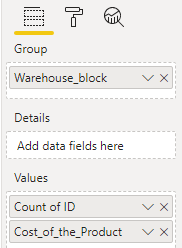


Рисунок 1.8 – Настройки визуализации «Treemap»

Результатом будет визуализация, где можно будет наблюдать общую стоимость продуктов, отправленных с каждого из складов компании. Визуализация «Treemap» представлена на рисунке 1.9.

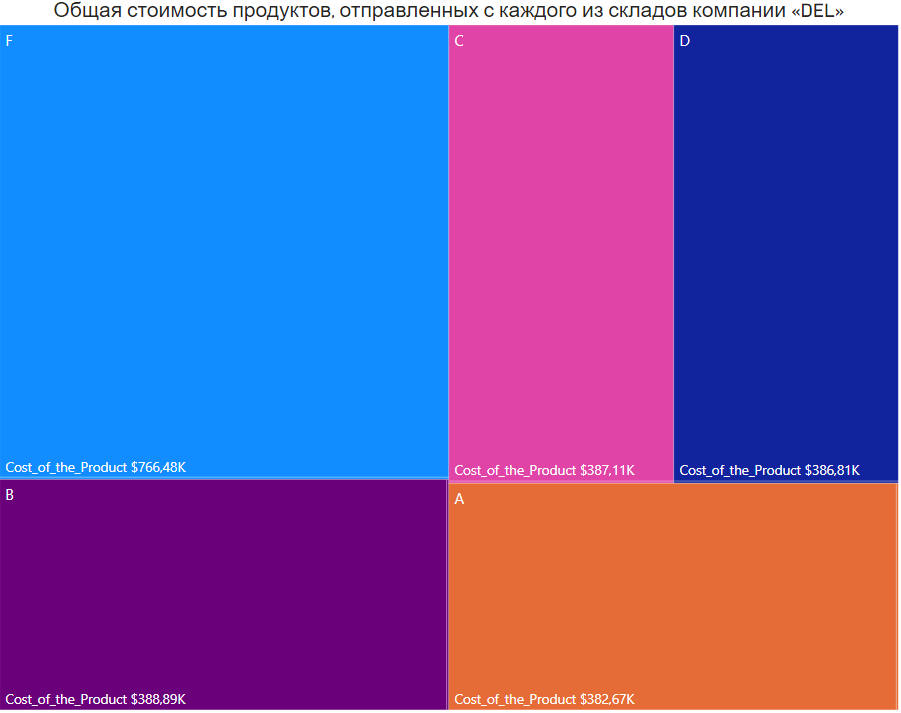


Рисунок 1.9 – Визуализация «Treemap» общей стоимости продуктов

Исходя из данной визуализации, можно сделать вывод, что по стоимости отправленных товаров лидирует склад «F».

Во вкладке «Format» были включены метки данных поля «Display data label options», а также заголовок был изменен и расположен по центру. В данной вкладке для эстетики могут быть настроены и другие элементы, такие как выбор цвета, размера, стиля шрифта и прочих настроек.

Далее рассмотрим визуализацию количества отправленных товаров с каждого склада компании «DEL» (рисунок 1.10).

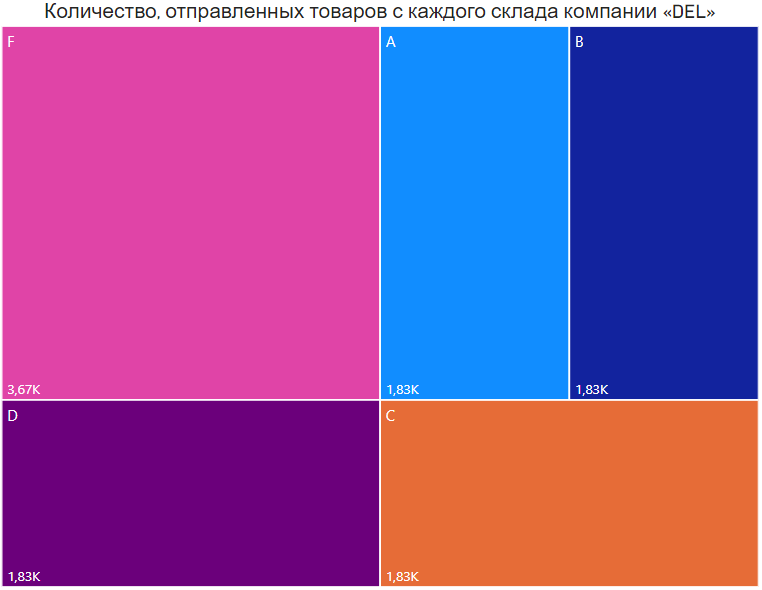


Рисунок 1.10 – Количество, отправленных товаров с каждого склада

Исходя из данной визуализации, можно сделать вывод, что наибольшее количество товаров отправляет склад «F».

Затем, исходя из анализа верхнего уровня, будем выдвигать гипотезы, которые в дальнейшем будут подтверждены или опровергнуты глубинным анализом в IBM SPSS Modeler. Проанализируем зависимость своевременности доставки товаров от рейтинга клиентов. Визуализация представлена на рисунке 1.11.

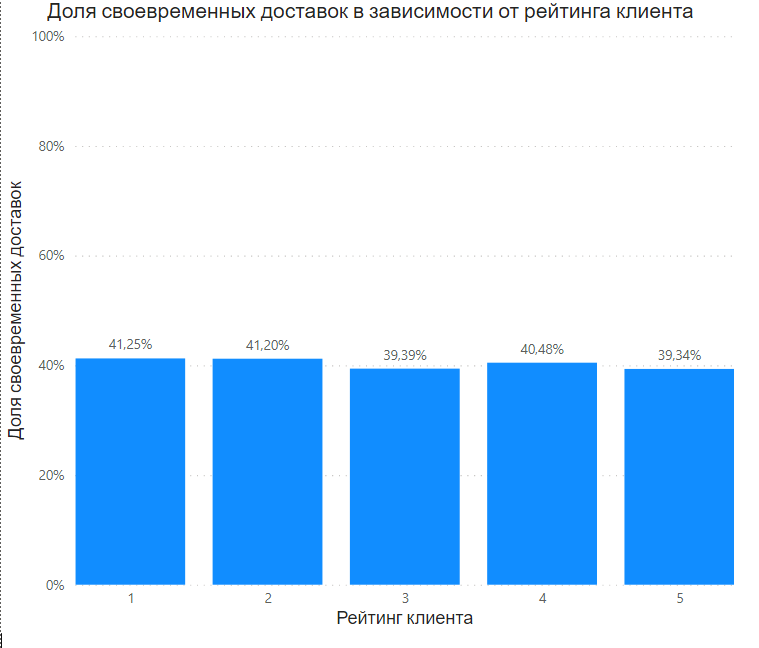


Рисунок 1.11 – Влияние рейтинга клиента на своевременность доставки

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – рейтинг клиента не влияет на своевременность доставки.

Проанализируем зависимость своевременности доставки товаров от важности продукта, классифицируемого компанией. Визуализация представлена на рисунке 1.12.



Рисунок 1.12 – Влияние важности товара на своевременность доставки

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – у товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки.

Проанализируем зависимость своевременности доставки от количества звонков клиентов по отгрузке товаров. Визуализация представлена на рисунке 1.13.

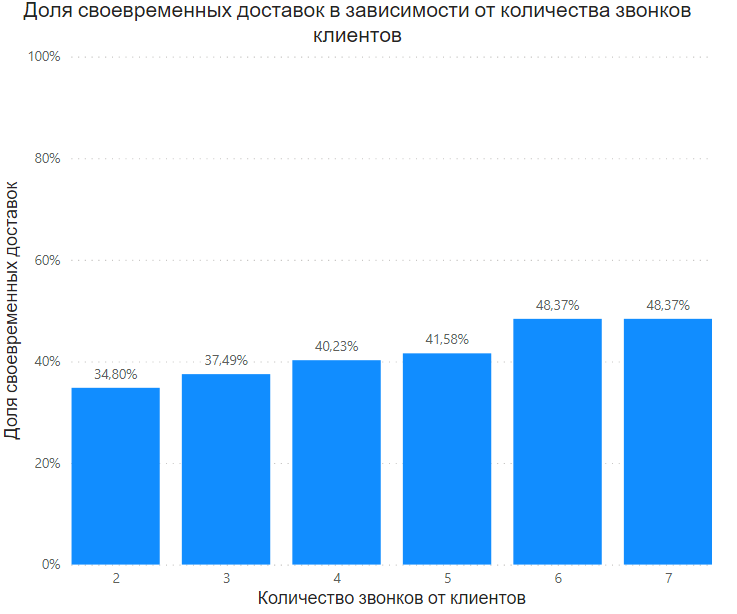


Рисунок 1.13 – Влияние количества звонков на своевременность доставки

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки.

Проанализируем зависимость своевременности доставки от количества предыдущих покупок. Визуализация представлена на рисунке 1.14.

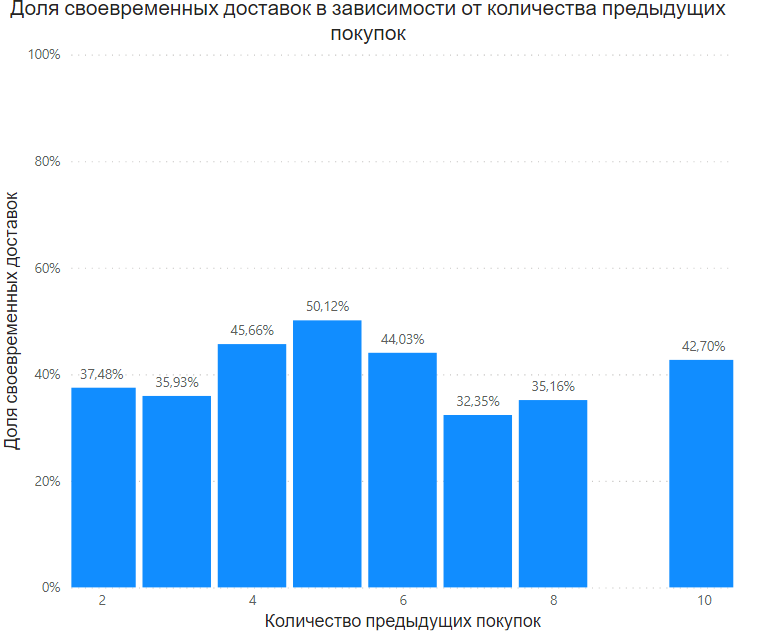


Рисунок 1.14 – Влияние предыдущих покупок на своевременность доставки

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – наличие предыдущих покупок не влияет на своевременную доставку.

Проанализируем зависимость своевременности доставки от средней стоимости товаров. Визуализация представлена на рисунке 1.15.

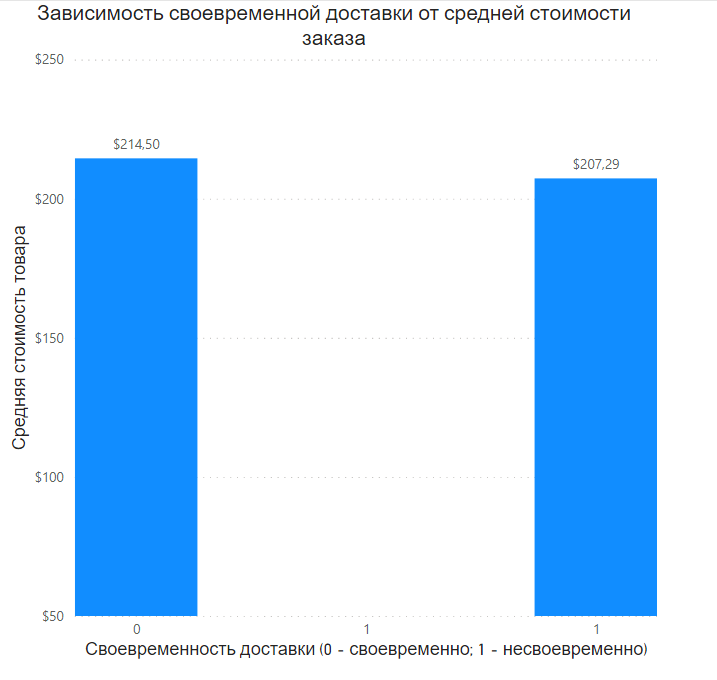


Рисунок 1.15 – Влияние средней стоимости товаров на своевременность доставки

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – стоимость заказа не влияет на своевременную доставку.

Проанализируем зависимость своевременности доставки от среднего веса товара. Визуализация представлена на рисунке 1.16.

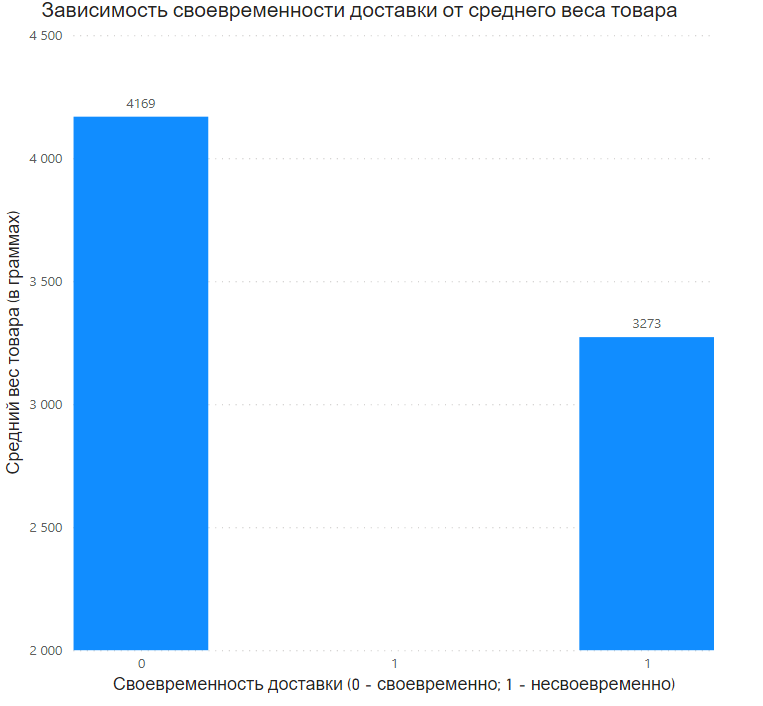


Рисунок 1.16 – Зависимость своевременности доставки от среднего веса товара

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – чем больше вес продукта, тем выше шанс на своевременную доставку.

Проанализируем зависимость своевременности доставки от величины скидки на товар. Визуализация представлена на рисунке 3.17.

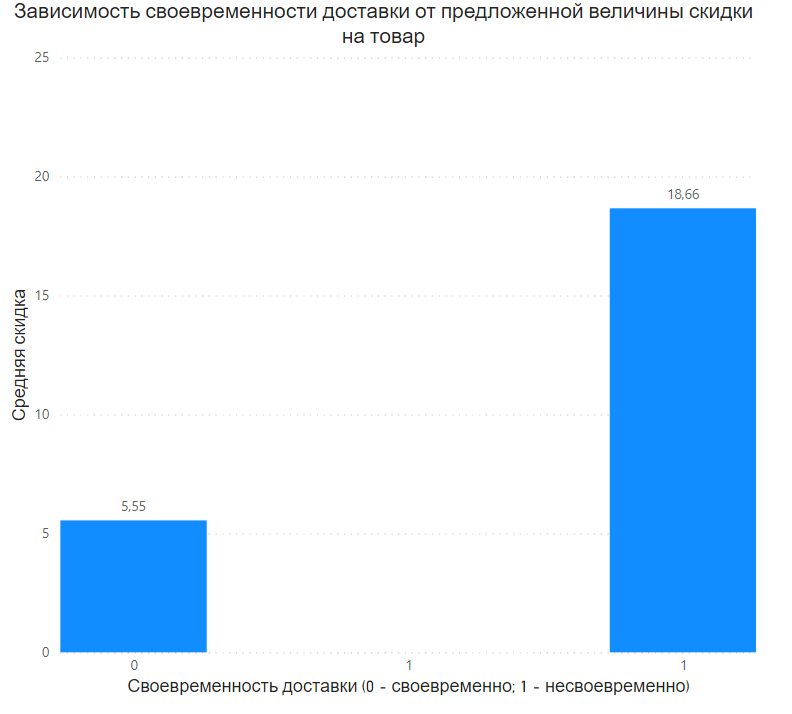


Рисунок 1.17 – Зависимость своевременности доставки от предложенной величины скидки на товар

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – чем выше скидка на товар, тем выше шанс несвоевременной доставки.

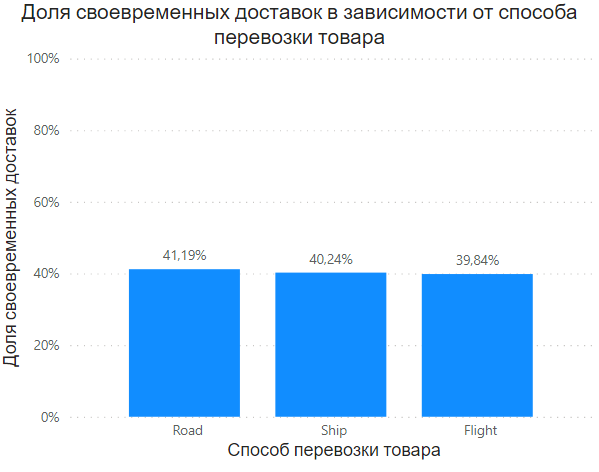
Проанализируем зависимость своевременности доставки от склада отправки товара клиенту. Визуализация представлена на рисунке 3.18.



Рисунок 1.18 – Зависимость своевременности доставки от склада отправки товара клиенту

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – склад отправки товара не влияет на своевременность доставки.

Проанализируем зависимость своевременности доставки от способа перевозки товара. Визуализация представлена на рисунке 1.19.

Рисунок 1.19 – Зависимость своевременности доставки от способа перевозки товара

Исходя из данной визуализации, можно выдвинуть следующую гипотезу – отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку.

## 1.4. Основные гипотезы исследования

В результате проведенного предварительного визуального анализа средствами Microsoft Power BI были выдвинуты следующие гипотезы:

1. Рейтинг клиента не влияет на своевременность доставки (Предположение подтверждения/опровержения гипотезы – будет отклонена).
2. У товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки (Предположение подтверждения/опровержения гипотезы – будет отклонена).
3. Чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки (Предположение подтверждения/опровержения гипотезы – будет подтверждена).
4. Наличие предыдущих покупок не влияет на своевременную доставку (Предположение подтверждения/опровержения гипотезы – будет подтверждена).
5. Стоимость заказа не влияет на своевременную доставку (Предположение подтверждения/опровержения гипотезы – будет подтверждена).
6. Чем больше вес продукта, тем выше шанс на своевременную доставку (Предположение подтверждения/опровержения гипотезы – будет отклонена).
7. Чем выше скидка на товар, тем выше шанс несвоевременной доставки (Предположение подтверждения/опровержения гипотезы – будет подтверждена).
8. Склад отправки товара не влияет на своевременность доставки (Предположение подтверждения/опровержения гипотезы – будет подтверждена).
9. Отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку (Предположение подтверждения/опровержения гипотезы – будет подтверждена).

В таблице 1.2 приведены гипотезы с рекомендациями для заказчика.

Таблица 1.2 – Сводная таблица рекомендаций для заказчика.

|  |  |  |
| --- | --- | --- |
| Гипотеза | Рекомендации при подтверждении гипотезы | Рекомендации при отвержении гипотезы |
| Рейтинг клиента не влияет на своевременность доставки. | Компании нет необходимости проводить реструктуризацию в политике ранжирования клиентов. | В целях сокращения числа неудовлетворенных клиентов необходимо придумать «прозрачную» систему ранжирования, а главное, обозначить преимущества в своевременности доставок в зависимости от рейтинга клиента. |
| У товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки. | Компании необходимо провести исследование для выявления факторов, влияющих на нестабильность доставок во времени. После этого проинвестировать в инструменты, позволяющие сократить влияние данных факторов. | Компании нет необходимости ставить в приоритет по своевременности доставку товаров с уровнем важности «High». |
| Чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки. | Необходимо устранить данную зависимость, так как своевременность доставки не должна зависеть от «настойчивости» клиентов. Это приводит к повышению негативного отношения к компании среди клиентов. | Отвержение гипотезы будет свидетельствовать об отсутствии необходимости в инвестициях с целью устранения эффекта от указанной зависимости. |
| Наличие предыдущих покупок не влияет на своевременную доставку. | Компании нет необходимости сосредотачивать усилия на устранении указанной зависимости. | Компании следует либо устранить эту зависимость, либо разработать систему, отражающую характер численного влияния числа и стоимости предыдущих покупок клиента на своевременность новых доставок. |
| Стоимость заказа не влияет на своевременную доставку. | Компании нет необходимости инвестировать в меры по устранению указанной зависимости. | Компании следует либо устранить эту зависимость, либо разработать систему, отражающую характер численного влияния стоимости заказа на своевременность его доставки. |
| Чем больше вес продукта, тем выше шанс на своевременную доставку. | Компании стоит обращать внимание на своевременную доставку товаров с небольшим весом. Также необходимо сосредоточить усилия на возможной компоновке легких по весу товаров. | Нет необходимости в инвестициях в целях поиска оптимального распределения товаров по весу в заказах клиентов. |
| Чем выше скидка на товар, тем выше шанс несвоевременной доставки. | Для товаров, отгружаемых со скидкой, стоит увеличить срок своевременной доставки, а также, возможно, изменить склады отправки товаров со скидкой. | Изменения в скидочной политике компании с целью увеличение процента своевременных доставок не требуются. |
| Склад отправки товара не влияет на своевременность доставки. | Компании нет необходимости изменять распределение товаров по складам при отправке. | Необходимо выделить наиболее «проблемные» склады и выделить возможные причины срыва срока доставок. |
| Отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку | Компании необходимо увеличить долю перевозок сухопутными путями. | Компании «DEL» не стоит сосредотачиваться на перевозках сухопутными путями. |

## 1.5. Требования к результатам

Функциональные требования:

* Достижение требуемого качества исследуемых данных, в данном случае, наличие пропущенных значений в датасете не более 1% от общего количества данных, удаление из данных аномальных значений;
* Обоснование выбора применяемых методов и инструментов предиктивного моделирования;
* Достижение требуемого качества предиктивной модели;
* Интерпретация результатов;
* Рекомендации по внедрению результатов и развитию предиктивной модели.

Специальные требования:

* Использование методологии IBM CRISP-DM для реализации проекта;
* Использование лидирующих платформ предиктивной аналитики для реализации проекта (IBM SPSS Modeler / IBM Watson Studio).

Требования к точности модели заявлены на уровне – 80%.

## 1.6. Методология, роли и этапы проведения исследования

В процессе подготовки кроме методологии CRISP-DM нами были рассмотрены и другие методологии, изучены результаты опросов на аналитических порталах [3]. Результаты представлены на рисунке 1.20.

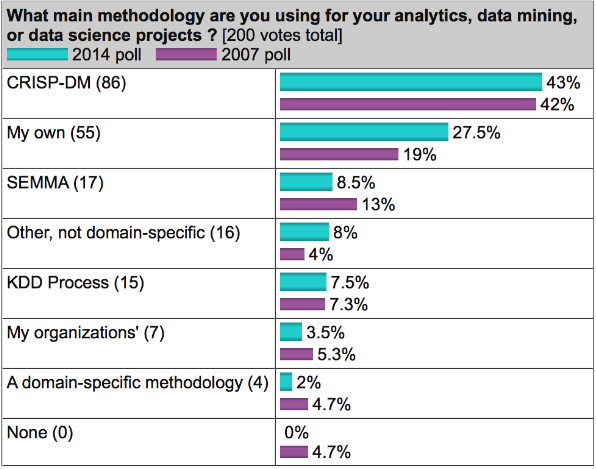


Рисунок 1.20. Данные опроса аналитического портала

В целом, методологии очень похожи друг на друга. Однако CRISP-DM заслужила популярность как наиболее полная и детальная. По сравнению с ней KDD является более общей и теоретической, а SEMMA – это просто организация функций по целевому предназначению в инструменте SAS Enterprise Miner и затрагивает исключительно технические аспекты моделирования, никак не касаясь бизнес-постановки задачи.

Поэтому после проведения анализа методологий и результатов опросов было принято решение, что данный проект будет реализован по методологии IBM CRISP-DM (Cross-Industry Standard Process for Data Mining), как на наиболее популярной, наиболее полной и детальной методологии.

CRISP-DM стандарт, описывающий общие процессы и подходы к аналитике данных, используемые в промышленных data-mining проектах независимо от конкретной задачи и индустрии.

Методология обеспечивает структурированный подход к интеллектуальному анализу данных, который поможет обеспечить успех нашему проекту.

Важным свойством методологии является уделение внимания бизнес-целям компании. Это позволяет руководству воспринимать проекты по анализу данных не как «песочницу» для экспериментов, а как полноценный элемент бизнес-процессов компании.

Вторая особенность – это довольно детальное документирование каждого шага. По мнению авторов, хорошо задокументированный процесс позволяет менеджменту лучше понимать суть проекта, а аналитикам – больше влиять на принятие решений.

Этапы методологии IBM CRISP-DM:

* Бизнес-анализ (Business understanding);
* Анализ данных (Data understanding);
* Подготовка данных (Data preparation);
* Моделирование (Modeling);
* Оценка качества модели (Evaluation);
* Развертывание (Deployment).

## 1.7. Выводы по результатам Этапа 1

В ходе выполнения первого этапа работы над проектом были выполнены следующие задачи:

* Дано краткое описание по проекту;
* Изучены данные для исследования;
* Исходя из анализа данных, сформулированы гипотезы исследования;
* Описаны функциональные и специальные требования к проекту;
* Рассмотрены различные методологии исследования;
* В качестве стандарта ведения проекта по интеллектуальному анализу данных была выбрана методология CRISP-DM;
* Определены задачи исследования и назначены ответственные по каждому этапу проекта.

# 2. АНАЛИЗ И ПОДГОТОВКА ДАННЫХ

## 2.1. Теоретический обзор приложений предиктивной аналитики

В ходе начала работы над проектом нами были самостоятельно рассмотрены приложения предиктивной аналитики, а также изучены специальные обзоры на такие инструменты как IBM SPSS Modeler и IBM Watson Studio [2]. Далее будут приведены возможности и преимущества каждой из платформ.

IBM SPSS Modeler является ведущим решением в сфере визуального анализа данных и машинного обучения.

Этим продуктом пользуются для подготовки и исследования данных, прогнозного анализа, управления моделями, внедрения моделей и машинного обучения с целью монетизации ресурсов данных.

Преимущества SPSS Modeler:

* Инструкции по пользованию в виде пошаговых этапов и примеров с объяснением;
* Инновационные решения с открытым кодом, включая R и Python;
* Повышение продуктивности экспертов по работе с данными любого профиля – от программистов до визуальных аналитиков;
* Применение гибридного подхода — в локальной среде и в общедоступном или частном облаке;
* IBM Cloud Pak for Data as a Service позволяет работать с SPSS Modeler в общедоступном облаке.

IBM Watson Studio – инструмент преимущественно для руководителей отделов, маркетологов и аналитиков, не имеющих глубоких познаний в статистике.

Возможности Watson Studio:

* Масса стандартных анализов и их в интерактивная визуализация;
* Автоматически анализирует качество загруженных данных и отображает рейтинг;
* Автоматически строит зависимости target переменной от одной, двух или комбинации полей, определяет predictive power каждой комбинации;
* Понимает запросы к датасету на естественном языке и выдает результаты.

Также был проанализирован рейтинг лучших платформ и инструментов для обработки данных [1].

Фрагмент рейтинга приведен на рисунке 2.1.

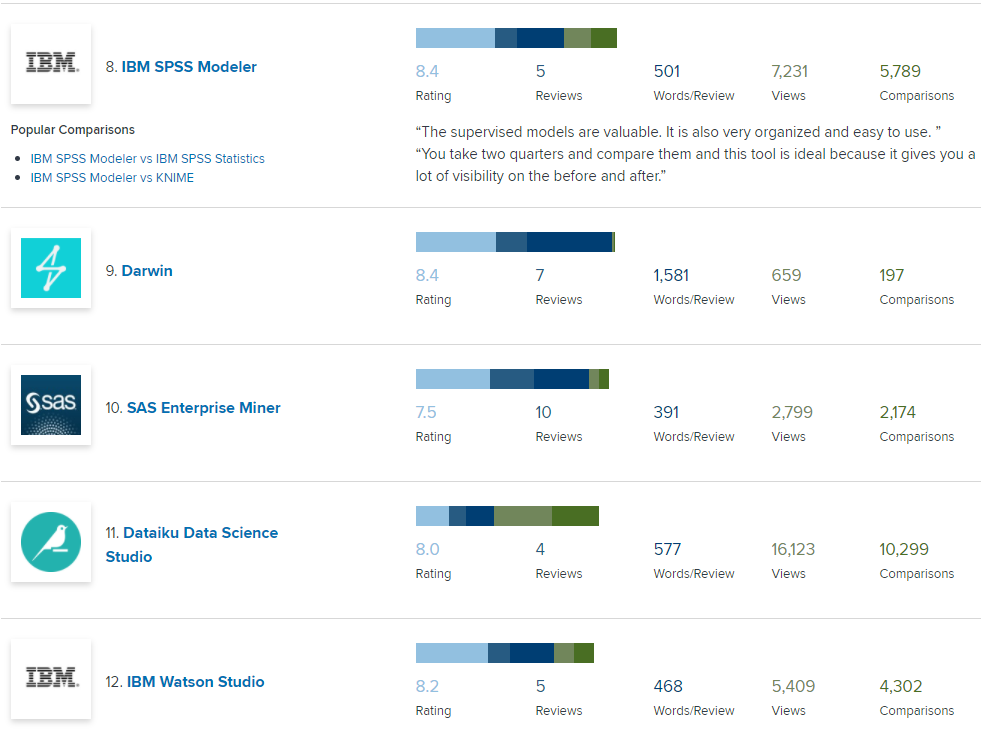


Рисунок 2.1 – Рейтинг лучших платформ и инструментов для обработки данных

Исходя из данных рейтинга, в котором Modeler занимает 8-ое место, а Watson 12-ое, а также наличия возможности в SPSS Modeler расположения данных локально, а не в облаке, как в Watson Studio, для реализации проекта нами был выбран IBM SPSS Modeler.

## 2.2. Планируемые методы исследования

Исходя из выбранной платформы, и так как целевая переменная представлена флаговым типом и обозначает факт своевременной доставки товара, то поставленная задача является задачей бинарной классификации.

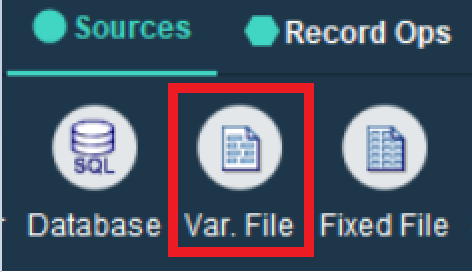
Для решения задачи бинарной классификации могут использоваться следующие методы (таблица 2.1):

Таблица 2.1 – Методы исследования

|  |  |
| --- | --- |
| Методы исследования | Описание метода |
| CHAID Tree | Метод основан на критерии хи-квадрат. Критерий хи квадрат является универсальным способом определения зависимости двух или более переменных, работающим для всех типов шкал, так как завязан на таблице сопряжённости, которую можно построить для любых переменных. |
| Quest Tree | Программа деревьев классификации, в которой используются улучшенные варианты метода рекурсивного квадратичного дискриминантного анализа и которая содержит ряд новых средств для повышения надежности и эффективности деревьев классификации, которые она строит. Используется для бинарной классификации. |
| C5.0 Tree | Программа деревьев классификации. Используется для бинарной классификации. |
| Random Forest | Это множество решающих деревьев. В задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо. |
| Logistic | Основная идея логистической регрессии для классификации состоит в том, чтобы установить формулу регрессии на граничной линии классификации в соответствии с существующими данными для классификации. |
| XGBoost Tree | XGBoost — алгоритм ML, основанный на дереве поиска решений и использующий фреймворк градиентного бустинга. |
| C&R Tree | C&R Tree напоминает «CHAID», но использует другой критерий для выращивания дерева. Используется для бинарной классификации. |
| Neural net | Математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей. |

## 2.3. Импортирование и первичный анализ данных

Для того чтобы импортировать файл с расширением «.csv» в IBM SPSS Modeler необходимо использовать ноду «Var. File» из палитры «Sources» (рисунок 2.2).

  
Рисунок 2.2 – Нода «Var.File»

В настройках данной ноды необходимо указать путь к файлу и выбрать разделитель (рисунок 2.3).

Изображение выглядит как текст

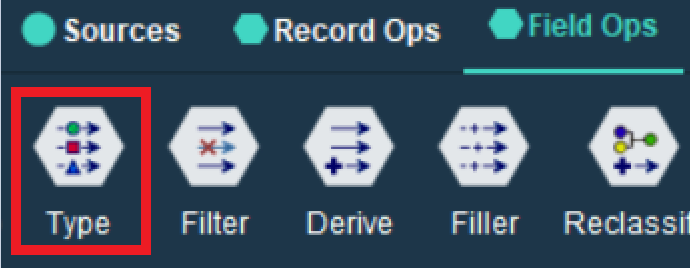
Автоматически созданное описание  
Рисунок 2.3 – Настройки для ноды «Var.File»

Далее необходимо определить тип каждого поля в соответствии с таблицей 2.2.

Таблица 2.2 – Тип полей.

|  |  |  |
| --- | --- | --- |
| Наименование поля | Содержание | Тип поля |
| ID | Уникальный номер поставки | Typeless (Тип отсутствует) |
| Warehouse\_block | Блоки склада | Nominal (Номинативный) |
| Mode\_of\_Shipment | Вид перевозки | Nominal (Номинативный) |
| Customer\_care\_calls | Кол-во звонков от клиента | Continuous (Непрерывный) |
| Customer\_rating | Рейтинг клиента | Ordinal (Упорядоченный) |
| Cost\_of\_the\_Product | Стоимость продукта | Continuous (Непрерывный) |
| Prior\_purchases | Количество предшествующих покупок | Continuous (Непрерывный) |
| Product\_importance | Классификация продукта по важности | Ordinal (Упорядоченный) |
| Gender | Пол | Flag (Флаговый) |
| Discount\_offered | Скидка на этот конкретный продукт | Continuous (Непрерывный) |
| Weight\_in\_gms | Вес в граммах | Continuous (Непрерывный) |
| Reached.on.Time\_Y.N | Целевая переменная | Flag (Флаговый) |

Для того чтобы назначить тип каждому полю, добавляем в наш поток ноду «Type» из палитры «Field Ops» (рисунок 2.4).

  
Рисунок 2.4 – Нода «Type»

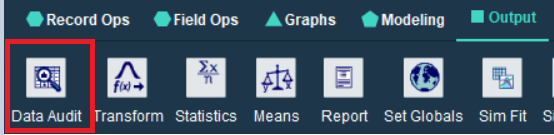
В настройках ноды «Type» указываем типы полей и нажимаем кнопку «Read values» (рисунок 2.5).

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 2.5 – Настройка типов полей

## 2.4. Оценка качества данных

После определения типа полей можно переходить к анализу качества импортируемых данных. Для этого воспользуемся нодой «Data audit» из палитры «Output» (рисунок 2.6).

  
Рисунок 2.6 – Нода «Data Audit»

Результат запуска ноды «Data audit» представлен на рисунке 2.7.

Изображение выглядит как стол

Автоматически созданное описаниеРисунок 2.7 – Результаты запуска ноды «Data Audit»

Перед построением модели обязательным является этап оценки качества имеющихся данных. 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 2.8.

Изображение выглядит как стол

Автоматически созданное описаниеРисунок 2.8 – Показатели качества для каждого поля

По результатам анализа качества исходных данных можно сделать следующий вывод: несмотря на присутствие пропущенных значений, качество датасета остаётся достаточно высоким для построения моделей. Однако стоит отметить, что некоторые модели требуют дополнительной обработки подобных значений.

Помимо пропущенных значений, данные могут содержать «выбросы» (аномалии), для выявления которых используют ноду «Anomaly» из палитры «Modeling» (рисунок 2.9).

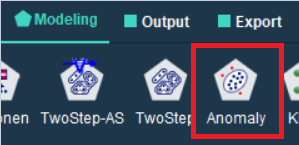
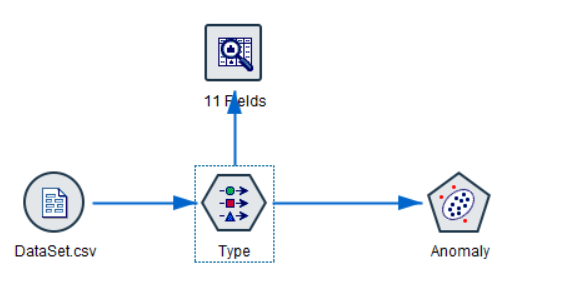
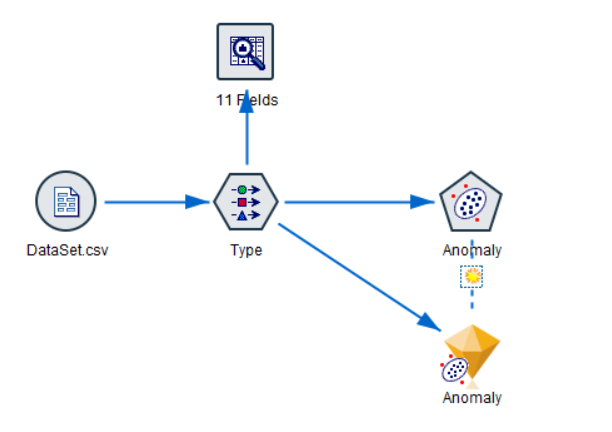
  
Рисунок 2.9 – Нода «Anomaly»

Схема потока представлена на рисунке 2.10.

  
Рисунок 2.10 – Схема потока на стадии анализа аномалий

В результате запуска модели создается «Слепок модели» желтого цвета в виде самородка («Nugget»), который представлен на рисунке 2.11.

  
Рисунок 2.11 – Схема потока после запуска ноды «Anomaly»

В настройках «Слепка модели» необходимо указать действия, которые мы хотим применить по отношению к аномальным данным. Для начала отобразим эти данные с помощью ноды «Table» (рисунок 2.12).

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 2.12 – Список записей с аномальными значениями

Количество записей, содержащих аномальные значения равно 109, удалим их из датасета с помощью опции «Discard records» в настройках слепка модели.

Таким образом, очищенный датасет теперь содержит 10 890 записей (рисунок 2.13).

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 2.13 – Удаление записей с аномальными значениями

## 2.5. Сценарии сокращения размерности задачи

Далее, с помощью ноды «Partition» из палитры «Field Ops» разобьём исходный датасет на тренировочную и тестовую выборки, где тестовая выборка составит 30% от общего количества записей (рисунок 2.14).

Изображение выглядит как текст

Автоматически созданное описание  
Рисунок 2.14 – Настройки разбиения датасета

После этого добавляем ноду «Filter» из палитры «Field Ops» для исключения полей, созданных в результате проведенного анализа аномалий (рисунок 2.15).

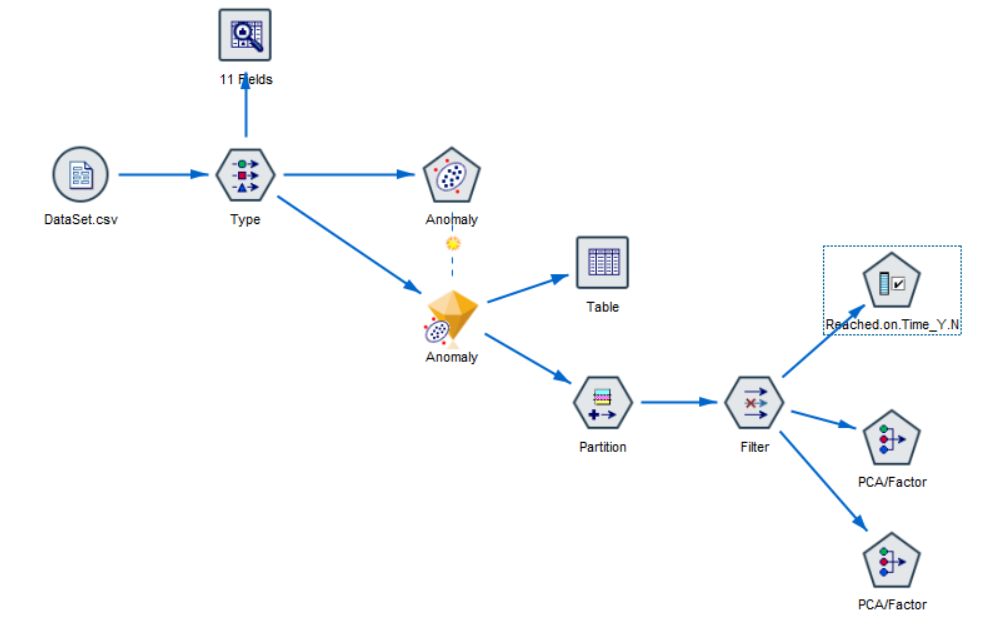
Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 2.15 – Применение фильтрации полей

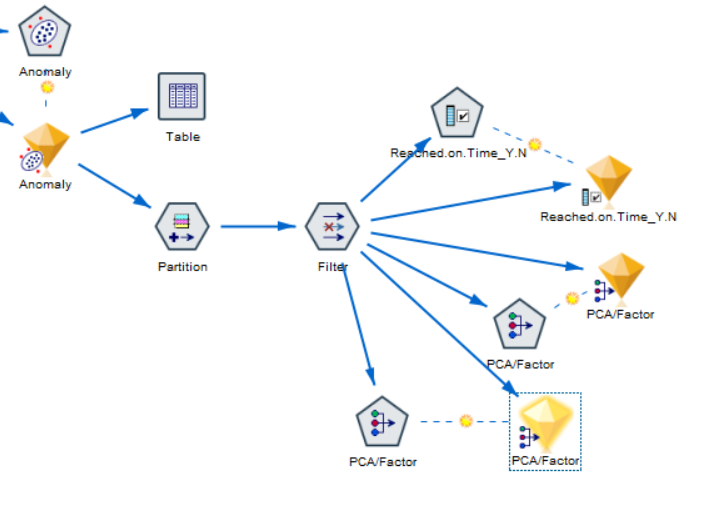
Далее, возникает три сценария развития потока:

1. Отбросить поля, слабо влияющие на целевую переменную, с помощью ноды «Feature selection» из палитры «Modeling».
2. Провести PCA-анализ для сокращения размерности задачи.
3. Провести Factor-анализ для сокращения размерности задачи.

Все три сценария представлены на рисунке 2.16.

  
Рисунок 2.16 – Сценарии для сокращения размерности задачи

В результате запуска сценариев потока появятся соответствующие «Слепки моделей» (рисунок 2.17).

  
Рисунок 2.17 – Результат запуска всех сценариев

## 2.6. Выводы по результатам Этапа 2.

В ходе выполнения этапа по подготовке и анализу данных были выполнены следующие задачи:

* Дан обзор инструментам исследования;
* Исходя из обзора, выбран инструмент исследования;
* Описаны и выбраны методы исследования;
* Импортированы исходные данные;
* Определены типы входных данных;
* Проведен первичный анализ данных;
* Проведена оценка качества данных, в ходе которой были удалены аномальные данные;
* Данные были разделены на две выборки – тренировочная и тестовая в соотношении 70/30;
* Проведен PCA-анализ для сокращения размерности задачи;
* Проведен Factor-анализ для сокращения размерности задачи.

# 3. МОДЕЛИРОВАНИЕ И ОЦЕНКА КАЧЕСТВА МОДЕЛИ

## 3.1. Настройка модели и описание применяемых методов прогнозирования

Следующим этапом исследования, в соответствии с методологией CRISP-DM, является этап моделирования. На данном этапе необходимо определить перечень возможных моделей, наиболее подходящих для решения поставленной задачи и учитывающих особенности имеющихся данных.

С целью определения подходящих моделей был проведен сравнительный анализ, результаты которого представлены в таблице 3.1.

Таблица 3.1 – Сравнительная характеристика моделей

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

Так как в имеющихся данных содержатся пропущенные значения, целевая переменная имеет флаговый тип, а предикторы являются как категорными так и непрерывными, то были отобраны следующие модели:

* QUEST
* C&R
* CHAID
* C5.0
* RandomForest

## 3.2. Описание сценариев и оценка качества модели

В качестве метрик для оценки качества бинарной классификации могут выступать следующие:

* Precision
* Recall
* F1-Score
* ROC (receiver operating characteristic)
* AUC (Area Under Curve)

Precision – количество истинно положительных исходов из всего набора положительных меток.

Recall – количество истинно положительных среди всех меток класса, которые были определены как «положительный».

F1-Score – Среднее гармоническое между Precision и Recall.

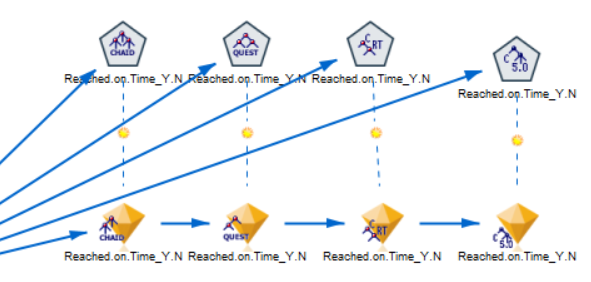
ROC – график, показывающий зависимость верно классифицируемых объектов положительного класса от ложно положительно классифицируемых объектов негативного класса.

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.

Так как компания «DEL» осуществляет множество критически важных поставок товаров, то для неё наиболее важным является – правильное определение своевременных доставок.

В связи с данными требованиями были выбраны следующие метрики качества – Accuracy, Recall, F1-Score, AUC.

Далее произведём построение моделей на исходных данных (рисунок 3.1).

  
Рисунок 3.1 – Использование решающих деревьев

С помощью ноды «Analysis» выведем показатели точности каждой модели. Показатели точности для CHAID дерева представлены на рисунке 3.2.

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.2 – Показатели точности для CHAID дерева

Метрики качества для QUEST дерева представлены на рисунке 3.3.

Изображение выглядит как стол

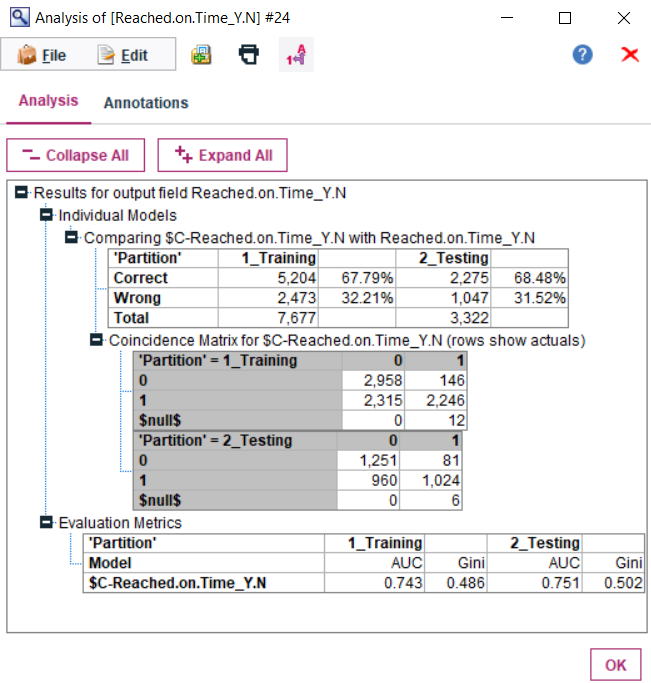
Автоматически созданное описание  
Рисунок 3.3 – Показатели точности для QUEST дерева

Метрики качества для C&R дерева показаны на рисунке 3.4.

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.4 – Показатели точности для С&R дерева

Метрики качества для С5.0 дерева отображены на рисунке 3.5.

  
Рисунок 3.5 – Показатели точности для С5.0 дерева

Таким образом, ни одна из моделей не соответствует заявленным требованиям качества.

Применим модель «RandomForest» к исходным данным (рисунок 3.6).

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.6 – Показатели точности для модели «Случайный лес»

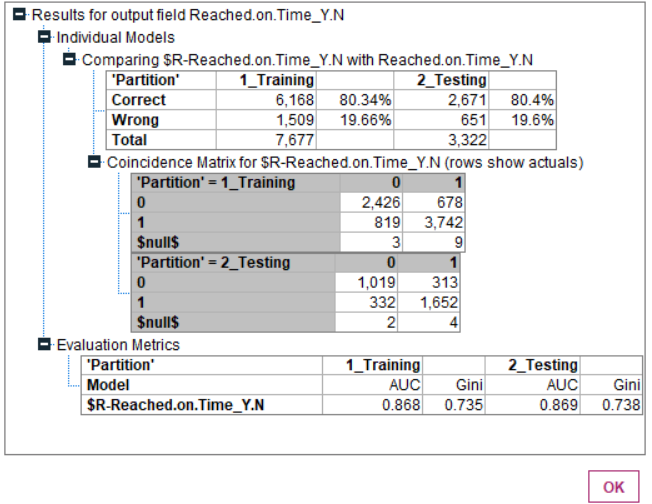
«Случайный лес» предоставляет высокую точность на тренировочной выборке, однако точность на тестовой выборке остаётся на низком уровне, это свидетельствует о переобучении алгоритма.

Для повышения точности моделей прибегнем к сценарию сокращения размерностей задачи. С помощью ноды «Feature Selection» предварительно сократим число предикторов (рисунок 3.7).

Изображение выглядит как текст, небо, карта

Автоматически созданное описание  
Рисунок 3.7 – Обучение моделей после сокращения размерности

Показатели точности для CHAID дерева представлены на рисунке 3.8.

  
Рисунок 3.8 – Показатели точности для CHAID дерева

Метрики качества для QUEST дерева представлены на рисунке 3.9.

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.9 – Показатели точности для QUEST дерева

Метрики качества для C&R дерева показаны на рисунке 3.10.

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.10 – Показатели точности для C&R дерева

Метрики качества для С5.0 дерева отображены на рисунке 3.11.

Изображение выглядит как стол

Автоматически созданное описание  
Рисунок 3.11 – Показатели точности для С5.0 дерева

Воспользуемся PCA-анализом для сокращения размерности задачи и применим перечисленные модели. Наилучшей моделью по показателям точности оказалась «CHAID» дерево (рисунок 3.12).

Изображение выглядит как стол

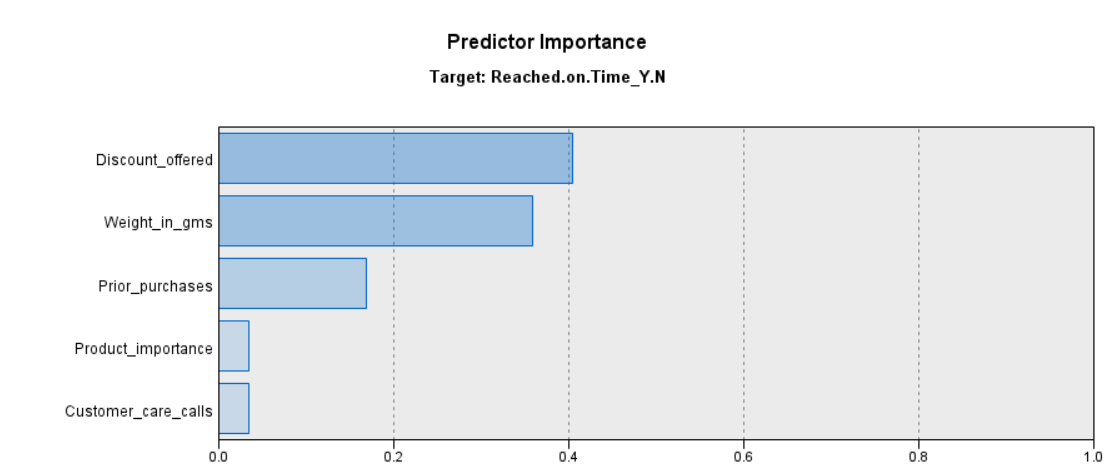
Автоматически созданное описание  
Рисунок 3.12 – Сокращение размерностей с помощью PCA-анализа

Однако, так как использование PCA-анализа приводит к усложнению интерпретируемости модели и повышает точность лишь на 2%, то было принято решение не прибегать к подобного вида средствам.

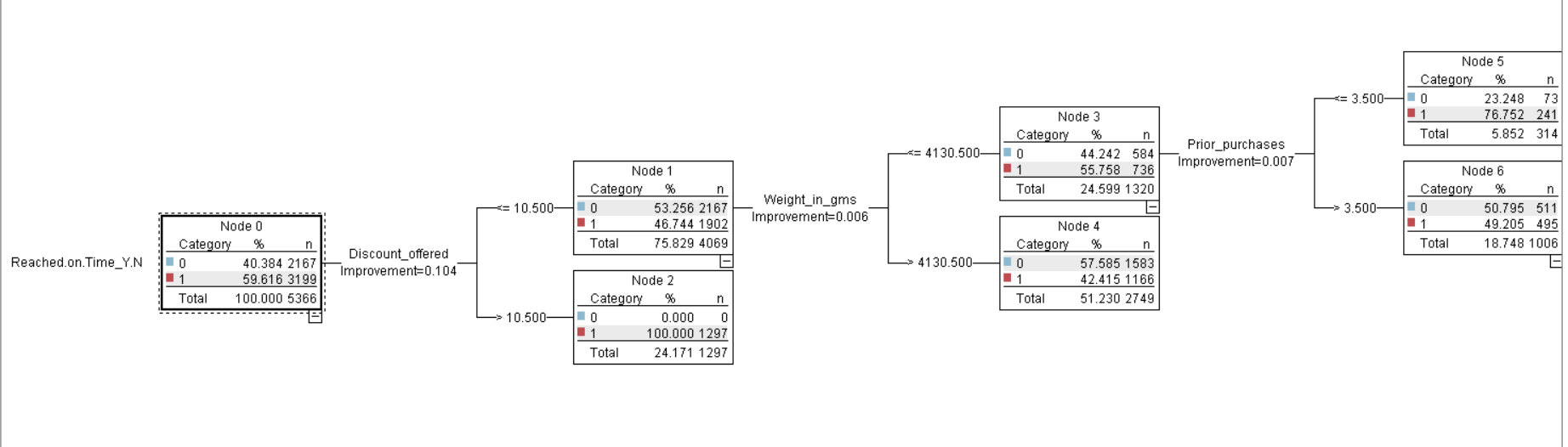
Таким образом, в качестве наилучшей модели было выбрано C&R дерево, метрики качества для которого следующие:

* Accuracy (Training) = 82,49%;
* Accuracy (Testing) = 82,75%;
* Recall (Training) = 0,76
* Recall (Testing) = 0,77.
* AUC (Training) = 0,862;
* AUC (Testing) = 0,861.

Важность предикторов представлена на рисунке 3.13.

Рисунок 3.13 – Степень важности предикторов

Построенное дерево решений представлено на рисунке 3.14.

Рисунок 3.14 – C&R дерево

## 3.3. Интерпретация полученных результатов

В таблице 3.2 Приведены общие сведения об использованных методах и метриках качества моделей.

Таблица 3.2 – Сводная таблиц использованных методов

|  |  |  |  |
| --- | --- | --- | --- |
| Метод | Accuracy | Recall | AUC |
| CHAID | 80.4% | 0.75 | 0.868 |
| QUEST | 82.3% | 0.776 | 0.862 |
| C&R | 82.75% | 0.77 | 0.862 |
| C5.0 | 82.63% | 0.77 | 0.86 |

После проведения глубинного анализа и построения прогнозных моделей с использованием разных методов прогнозирования были получены следующие результаты:

1. Подтверждена гипотеза, о том, что рейтинг клиента не влияет на своевременность доставки.
2. Отклоняется гипотеза о том, что у товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки.
3. Подтверждена гипотеза, о том, что чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки.
4. Отклоняется гипотеза о том, что наличие предыдущих покупок не влияет на своевременную доставку.
5. Отклоняется гипотеза о том, что стоимость заказа не влияет на своевременную доставку.
6. Подтверждена гипотеза о том, что чем больше вес продукта, тем выше шанс на своевременную доставку.
7. Подтверждена гипотеза о том, что чем выше скидка на товар, тем выше шанс несвоевременной доставки.
8. Подтверждена гипотеза о том, что склад отправки товара не влияет на своевременность доставки.
9. Отклоняется гипотеза о том, что отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку

## 3.4. Выводы по результатам Этапа 3

В ходе выполнения третьего этапа работы над проектом были выполнены следующие задачи:

* Описаны и выбраны применяемые методы прогнозирования;
* Настроена модель для прогнозирования;
* Продемонстрировано построение моделей с использованием разных методов прогнозирования;
* Определены наиболее важные предикторы, влияющие на своевременность доставки;
* Применен сценарий сокращения размерностей задачи;
* Осуществлена проверка качества моделей.

# 4. ОЦЕНКА РЕЗУЛЬТАТОВ ПРОЕКТА

## 4.1. Описание результатов проекта

В ходе работы над проектом участники проектной команды выполнили стоящие перед ними задачи, результатом проекта является разработанная и готовая к применению модель прогнозирования своевременности доставки товаров клиентам.

На этапе моделирования и оценки качества модели в качестве наилучшей модели было выбрано C&R дерево, имеющего следующие метрики качества:

* Accuracy (Training) = 82,49%;
* Accuracy (Testing) = 82,75%;
* Recall (Training) = 0,76
* Recall (Testing) = 0,77.
* AUC (Training) = 0,862;
* AUC (Testing) = 0,861.

Далее представлены наиболее важные предикторы наилучшей модели (от наиболее важного предиктора к наименее важному предиктору):

* Величина скидки на товар;
* Вес товара в граммах;
* Наличие предыдущих покупок;
* Степень важности продукта;
* Количество звонков от покупателей по отгрузке товаров.

Исходя из моделирования на более глубоком уровне, с применением платформы IBM SPSS Modeler, были исследованы и подтверждены/отклонены гипотезы таблицы 4.1.

Таблица 4.1 – Сведения о гипотезах

|  |  |
| --- | --- |
| Гипотеза | Предположение о подтверждении / отклонении гипотезы |
| 1.Рейтинг клиента не влияет на своевременность доставки. | Подтверждается |
| 2.У товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки. | Отклоняется |
| 3.Чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки. | Подтверждается |
| 4.Наличие предыдущих покупок не влияет на своевременную доставку. | Отклоняется |
| 5.Стоимость заказа не влияет на своевременную доставку. | Отклоняется |
| 6.Чем больше вес продукта, тем выше шанс на своевременную доставку. | Подтверждается |
| 7.Чем выше скидка на товар, тем выше шанс несвоевременной доставки. | Подтверждается |
| 8.Склад отправки товара не влияет на своевременность доставки. | Подтверждается |
| 9.Отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку. | Отклоняется |

## 4.2. Описание сформированных компетенций

В ходе работы над проектом у участников проектной команды были сформированы и усовершенствованы следующие компетенции, находящиеся в таблице 4.2.

Таблица 4.2 – Сформированные компетенции.

|  |  |
| --- | --- |
| Коммуникативные | Подготовлен план работы над проектом. |
| Назначены ответственные лица по выполнению этапов проекта. |
| Подготовлен план встреч по обсуждению работы над проектом. |
| Проведены встречи по работе над проектом. |
| Усовершенствованы навыки работы в команде. |
| Аналитические | Проанализирована методология CRISP-DM. |
| Проанализированы различные источники данных. |
| Определен датасет для решения задач проекта |
| Описана постановка бизнес-задачи. |
| Исходя из анализа данных, сформулированы гипотезы. |
| Проведена оценка и интерпретация результатов построения модели. |
| Hard-skills | Приобретены навыки работы с Power BI. |
| Приобретены навыки работы с IBM SPSS Modeler. |
| Проведены работы по очистке, анализу и оценки качества данных в IBM SPSS Modeler. |
| Усовершенствованы навыки построения моделей. |

## 4.3. Оценка перспектив развития полученных решений

Выявленные факторы, наиболее значимо влияющие на своевременность доставки, могут быть проанализированы и учтены при регистрации и указании даты своевременного прибытия товара, а также для повышения доли своевременных доставок.

Использование разработанной модели позволит компании «DEL» прогнозировать своевременность доставки для клиентов и в случае предсказания несвоевременности увеличить срок доставки.

В целях улучшения прогнозирования стоит использовать обогащение данных. Например, улучшить качество прогнозирования смогут данные в виде следующих сведений:

* Дата поставки товара;
* Дата прибытия товара;
* Сведения о метеорологических явлениях по маршруту следования доставки товара;
* Сведения об авариях по маршруту следования доставки товара;
* Сведения о загруженности транспортных маршрутов.

# ЗАКЛЮЧЕНИЕ

Итогом проекта является разработанная модель предиктивной аналитики, предназначенная для выявления факторов, наиболее сильно влияющих на своевременность доставки.

В качестве наилучшей модели было выбрано C&R дерево, имеющее следующие метрики качества:

* Accuracy (Training) = 82,49%;
* Accuracy (Testing) = 82,75%;
* Recall (Training) = 0,76
* Recall (Testing) = 0,77.
* AUC (Training) = 0,862;
* AUC (Testing) = 0,861.

Определение данных факторов и оценки их влияния позволит компании «DEL» принять меры для снижения количества несвоевременных доставок.

Наиболее важные факторы модели приведены ниже:

* Величина скидки на товар;
* Вес товара в граммах;
* Наличие предыдущих покупок;
* Степень важности продукта;
* Количество звонков от покупателей по отгрузке товаров.

По завершению исследования участниками проектной команды были решены следующие основные задачи:

* Описана бизнес-задача;
* Проведен верхнеуровневый анализ данных в MS Power BI;
* Сформулированы гипотезы исследования;
* Построены модели с использованием разных методов прогнозирования;
* Описаны сценарии моделей;
* Оценено качество модели;
* Проинтерпретированы полученные результаты исследования;
* Описаны результаты исследования в форме отчета.

# СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Best Data Science Platforms & Tools [Электронный ресурс]. Режим доступа: <https://www.itcentralstation.com/categories/data-science-platforms>. Дата обращения: 23.11.2021.
2. Compare IBM SPSS Modeler vs. IBM Watson Studio [Электронный ресурс]. Режим доступа: <https://www.itcentralstation.com/products/comparisons/ibm-spss-modeler_vs_ibm-watson-studio>. Дата обращения: 24.11.2021.
3. CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Электронный ресурс]. Режим доступа: <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Дата обращения: 24.11.2021.
4. CRISP-DM: Проверенная методология [Электронный ресурс]. Режим доступа: <https://habr.com/ru/company/lanit/blog/328858/>. Дата обращения: 26.11.2021.
5. E-Commerce Shipping Data [Электронный ресурс]. Режим доступа: <https://www.kaggle.com/prachi13/customer-analytics>. Дата обращения: 25.11.2021.
6. Визуализация данных [Электронный ресурс]. Режим доступа: <https://powerbi.microsoft.com/ru-ru/>. Дата обращения: 25.11.2021.
7. Машинное обучение [Электронный ресурс]. Режим доступа: <https://www.ibm.com/ru-ru/cloud/learn/machine-learning#toc----UhBWj8uU>. Дата обращения: 25.11.2021.
8. Машинное обучение: Методы и способы [Электронный ресурс]. Режим доступа: <https://www.osp.ru/cio/2018/05/13054535>. Дата обращения: 25.11.2021.
9. Оценка моделей ML/DL: матрица ошибок, Accuracy, Precision и Recall [Электронный ресурс]. Режим доступа:<https://pythonru.com/baza-znanij/metriki-accuracy-precision-i-recall>. Дата обращения: 26.11.2021.
10. Подготовка данных для машинного обучения [Электронный ресурс]. Режим доступа: <https://docs.microsoft.com/ru-ru/azure/architecture/data-science-process/prepare-data>. Дата обращения: 26.11.2021.