NATIONAL RESEARCH UNIVERSITY

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

Руководитель проекта:
Брускин Сергей Наумович,
доцент департамента бизнес-информатики, к.э.н.

Выполнил студент
группы МБИ-202:
Андросов Алексей

Moscow, 2021

Описание бизнес-кейса

Кейс: Компания «DEL» является австралийской компанией, которая занимается грузоперевозками как внутри страны, так и за ее пределами.

Компания использует следующие способы доставки товаров:

- авиаперевозки;
- морские перевозки;
- сухопутные перевозки.

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

Проблема: множество жалоб от клиентов из-за несвоевременности доставки.

Возможное решение: исследование и поиск закономерностей в данных для выявления факторов влияющих на своевременность доставки с помощью внедрения разработанной модели предиктивной аналитики.

Цель и задачи исследования

Цель: разработка и внедрение модели прогнозирования своевременности доставки товаров клиентам крупной транспортно-логистической компании.

Объект исследования: компания в сфере грузоперевозок – «DEL».

Предмет исследования: факторы, влияющие на своевременность доставки товаров компании «DEL».

Основные задачи исследования:

- Описать бизнес-задачу проекта;
- Провести предварительный анализ с использованием MS Power BI;
- Сформулировать гипотезы исследования;
- Разработать модель для предсказания своевременности доставки товаров компьютерной периферии клиентам;
- Описать используемые методы прогнозирования и сценарии моделирования;
- Оценить качество модели;
- Проинтерпретировать полученные результаты исследования.

Описание данных

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

В наборе данных 11000 записей и 12 полей. Факторы влияющие на своевременность доставки:

- Вид перевозки товара;
- Количество звонков от клиента;
- Рейтинг клиента;
- Стоимость товара;
- Вес товара;
- Скидка на товар и другие.

Целевая переменная: факт своевременности доставки товара.

About this file

Contains Cleaned 10999 observations of 12 variables.

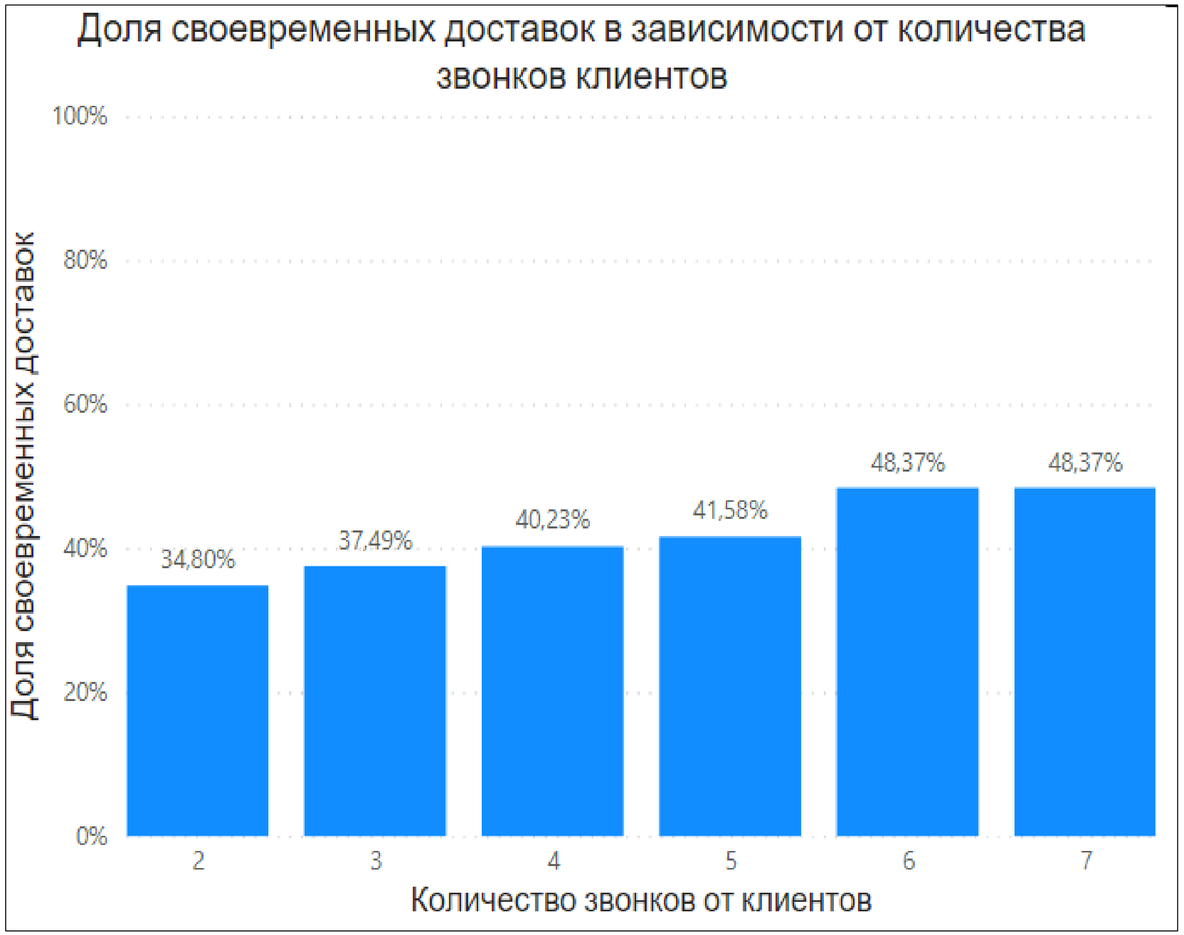
Warehouse_block	Mode_of_Shipment	Customer_care_c...	Customer_rating	Cost_of_the_Prod...
The Company have big Warehouse which is divided in to block such as A,B,C,D,E.	The Company Ships the products in multiple way such as Ship, Flight and Road.	The number of calls made from enquiry for enquiry of the shipment.	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best)	Cost of the Product in US Dollars.
F 33% D 17% Other (5499) 50%	Ship 68% Flight 16% Other (1760) 16%	2 7	1 5	96 310
D	Flight	4	2	177
F	Flight	4	5	216
A	Flight	2	2	183
B	Flight	3	3	176
C	Flight	2	2	184
F	Flight	3	1	162
D	Flight	3	4	250
F	Flight	4	1	233
A	Flight	3	4	150
B	Flight	3	2	164

Методология и инструменты исследования

Методология/Продукт	Наименование	Предназначение
CRISP-DM	CRISP-DM	Методология, выступающая в качестве стандарта ведения проекта по интеллектуальному анализу данных
Power BI	MS Power BI	Проведение предварительного визуального анализа для выдвижения возможных гипотез о зависимостях в данных
IBM SPSS MODELER	IBM SPSS Modeler	Подготовка данных, построение моделей и оценка точности проведенного моделирования

Ссылки на обзор приложений предиктивной аналитики и рейтинг лучших платформ:
https://www.itcentralstation.com/products/comparisons/ibm-spss-modeler_vs_ibm-watson-studio.
https://www.itcentralstation.com/categories/data-science-platforms

Предварительный анализ средствами MS Power BI



Чем выше скидка на товар, тем выше шанс несвоевременной доставки

Чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки

Гипотезы исследования

Основываясь на результатах проведенного предварительного анализа с использованием MS Power BI были выдвинуты следующие гипотезы:

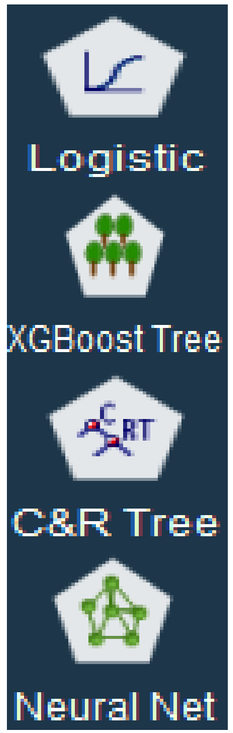
1. Рейтинг клиента не влияет на своевременность доставки.
2. У товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки.
3. Чем больше звонков совершено клиентом по отгрузке заказа, тем выше вероятность своевременной доставки.
4. Наличие предыдущих покупок не влияет на своевременную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

Методы исследования

Исходя из того, что 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:

- CHAID Tree
- Quest Tree
- C5.0 Tree
- Random Forest

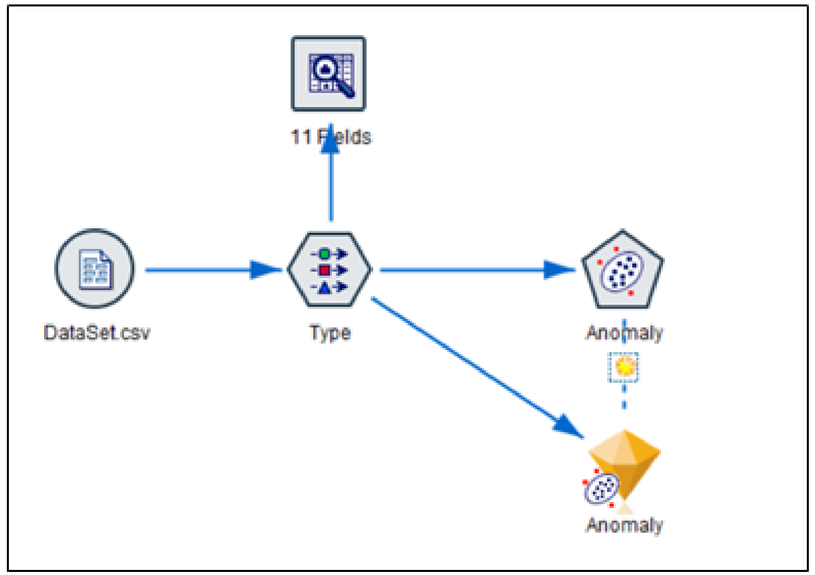


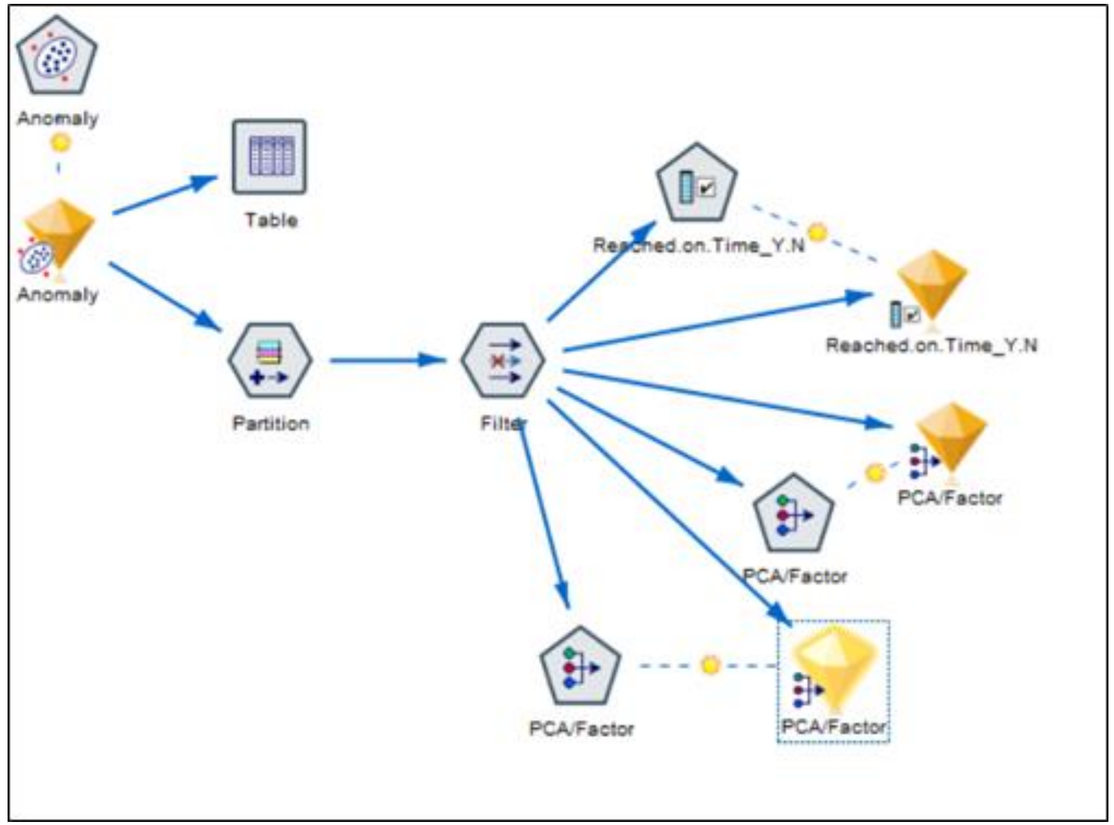
- Logistic
- XGBoost Tree
- C&R Tree
- Neural net

Анализ, подготовка и оценка качества данных

В ходе данных этапов были проведены следующие действия:

- Проанализированы количество пропущенных значений, бланковых значений и пустых строк по каждому полю;
- Удалены аномалии в данных (109 записей);
- Проведен PCA-анализ.
- Проведен Factor-анализ.

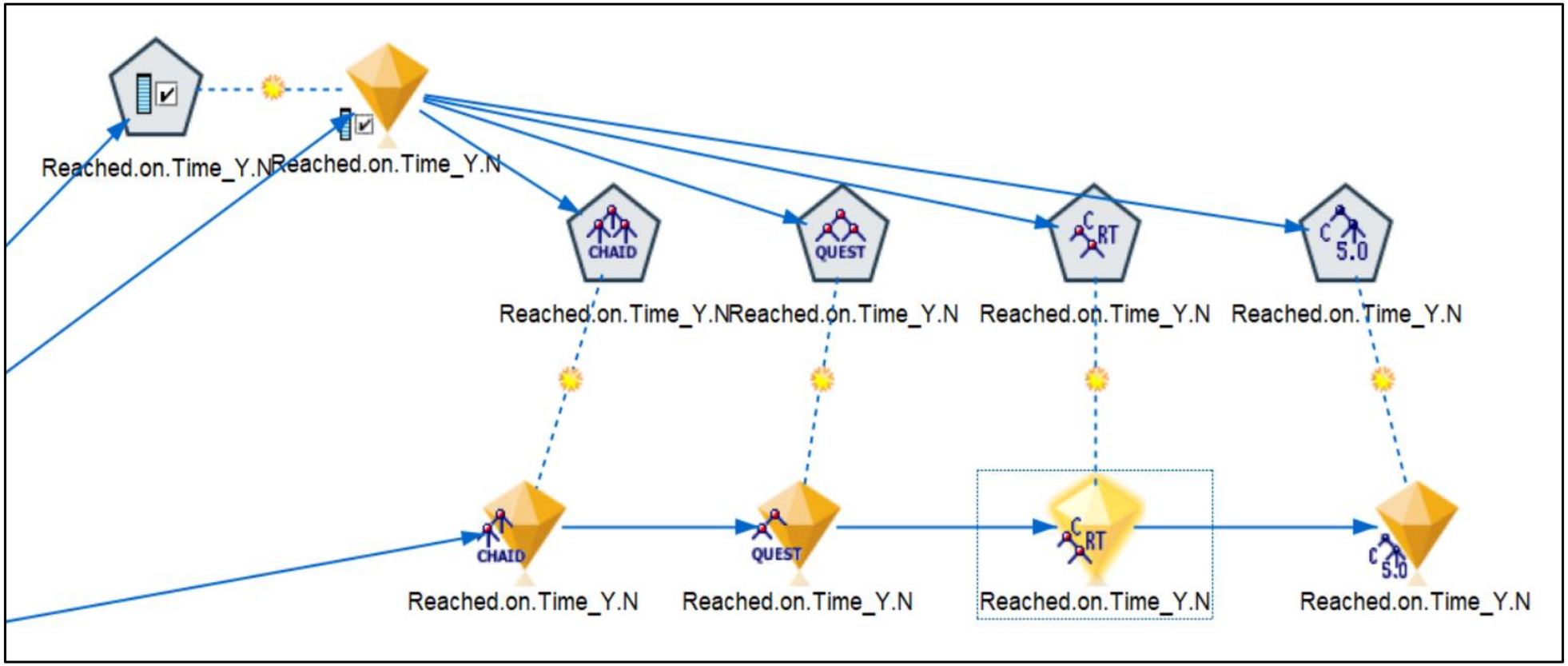




Моделирование и оценка качества моделирования

Для моделирования были использованы следующие методы:

- Quest Tree
- C&R Tree
- CHAID Tree
- C5.0
- Random Forest



Моделирование и оценка качества моделирования

Для оценки качества моделей были выбраны следующие метрики качества:

- Accuracy
- Recall
- AUC

Показатели точности построенного C&R дерева представлены на рисунке слева. Сводная таблица используемых методов и оценки качества моделей представлена на рисунке справа.

Results for output field Reached.on.Time_Y.N

Individual Models

Comparing $R-Reached.on.Time_Y.N with Reached.on.Time_Y.N

'Partition'	1_Training		2_Testing	
Correct	6,333	82.49%	2,749	82.75%
Wrong	1,344	17.51%	573	17.25%
Total	7,677		3,322	

Coincidence Matrix for $R-Reached.on.Time_Y.N (rows show actuals)

'Partition' = 1_Training	0	1
0	2,587	517
1	815	3,746
$null$	3	9

'Partition' = 2_Testing	0	1
0	1,094	238
1	329	1,655
$null$	2	4

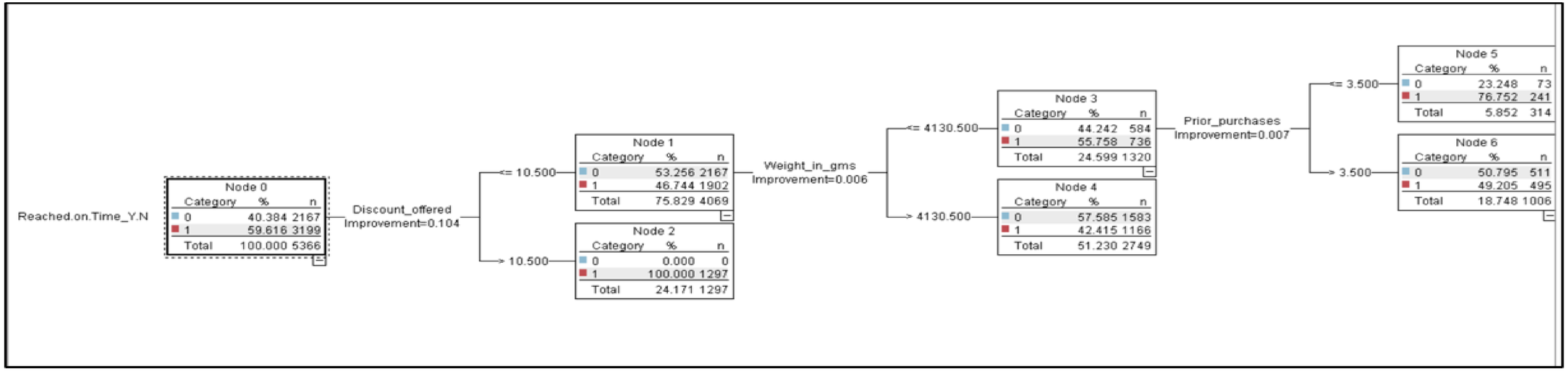
Evaluation Metrics

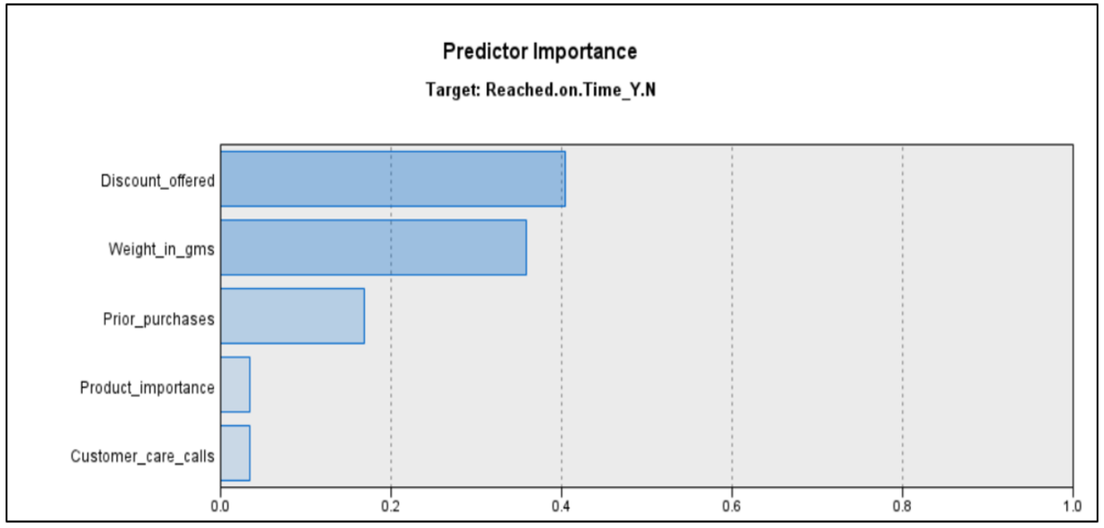
'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-Reached.on.Time_Y.N	0.862	0.724	0.861	0.723

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

Моделирование и оценка качества моделирования

В качестве наилучшей модели было выбрано C&R дерево. Схема представлена на рисунке ниже.





Наиболее важные предикторы C&R дерева представлены на рисунке слева:

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

Гипотезы и рекомендации для заказчика (1/2)

Гипотеза	Рекомендация
Рейтинг клиента не влияет на своевременность доставки.	Компании нет необходимости проводить реструктуризацию в политике ранжирования клиентов.
У товара с уровнем важности продукта «High», выше вероятность несвоевременной доставки.	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.
Наличие предыдущих покупок не влияет на своевременную доставку.	Компании следует либо устранить эту зависимость, либо разработать систему, отражающую характер численного влияния числа и стоимости предыдущих покупок клиента на своевременность новых доставок.

Гипотезы и рекомендации для заказчика (2/2)

Гипотеза	Рекомендация
Стоимость заказа не влияет на своевременную доставку.	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.
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.
Склад отправки товара не влияет на своевременность доставки.	Компании нет необходимости изменять распределение товаров по складам при отправке.
Отправка товаров сухопутными перевозками увеличивает шансы на своевременную доставку	Компании «DEL» не стоит сосредотачиваться на перевозках сухопутными путями.

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.

Выводы по проекту

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, имеющего следующие метрики качества:

- Accuracy = 82,75%.
- Recall = 0,77.
- AUC = 0,861.

Наиболее важные предикторы наилучшей модели (от наиболее важного 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.