ФИО

ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА

КЕЙС №1

1. https://colab.research.google.com/drive/1-L6dKXGGseOMoOeAVNqihl-weXQkaPJO?usp=sharing
2. https://drive.google.com/file/d/1Y7g_VspOdUkW4r3wxx5y_Ur2muArl7j5/view?usp=sharing (Loginom)
3. https://drive.google.com/file/d/17HP1LGPBLDrXNX69-nBsvRpJ4mqe6N_3/view?usp=sharing (Knime)

1. Цели и задачи.

В кейсе №1 содержится 4 таблицы в Excel с информацией о клиентах и продажах магазина велосипедов и сопутствующих товаров за период с 2014 по 2016 года. Магазин занимается продажами велосипедов и аксессуаров в США, Австралии, Германии, Франции, Великобритании. Цель работы предсказать методами машинного обучения возможные факторы, влияющие на отток покупателей. Для решения поставленной задачи будут сформированы и обучены модели. Решение задачи относится к классу бинарной классификации.

2. Описание датасета.

Датасет кейса №1 состоит из 4 таблицы в Excel – Customer, Product, Sales, Territories.

Sales – 58189 строк, 13 столбцов (основная таблица) – данные о продажах товаров по датам в разрезе заказов по каждому клиенту и отдельно по строкам в заказах, количеству товаров, цене, сумме, налогу, номеру заказа, дате поставки.

Customer – 18484 строки, 21 столбец – данные по клиентам магазина, включающие информацию по полу, возрасту, образованию, роду занятий, наличию детей, из них находящихся дома, наличию автомобилей, наличию дома, годового дохода, расстоянию до работы, дате первого заказа, адреса, страны и города проживания.

Product – 606 строк, 13 столбцов – данные о товарах, наименовании, коде наименования, модели, линии, категории, подкатегории, цене, ссылке на изображение.

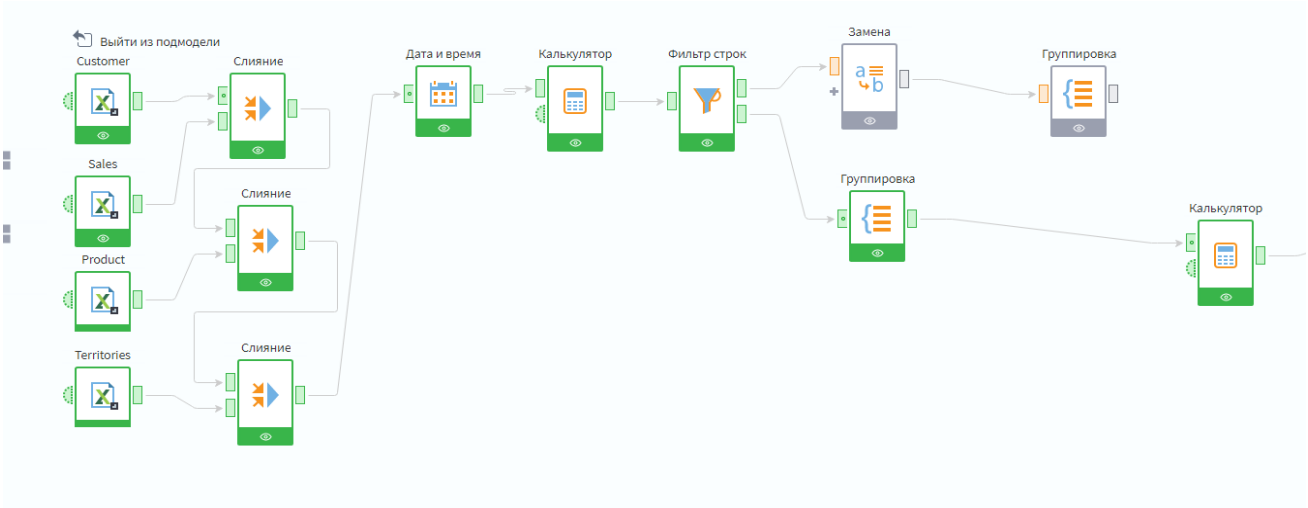
Territories – 11 строк, 6 столбцов – данные о коде территории, стране, группе.

3. Объединение и очистка данных

Объединение и очистка датасета проведена в платформе Loginom.

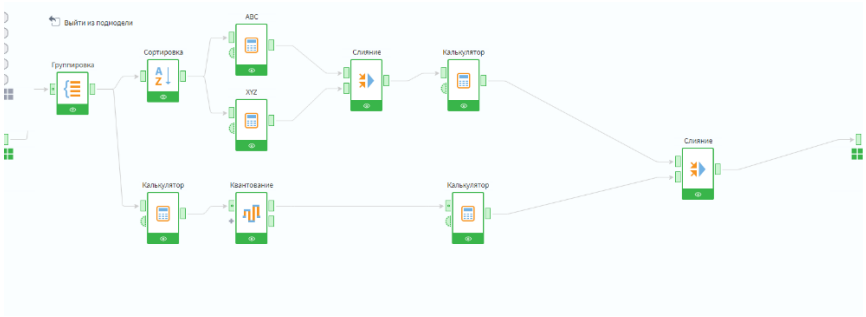
После объединения в итоговой общей таблице – 58755 строк, 40 столбцов. 566 клиентов в период с 2014 по 2016 года не покупали товары в данном магазине, соответственно по ним не было информации в таблице Sales. Строки по данным клиентам удалены. В итоге

в рабочем варианте чистых данных 58189 строк, 38 столбцов. Добавлена колонка Age, Profit.

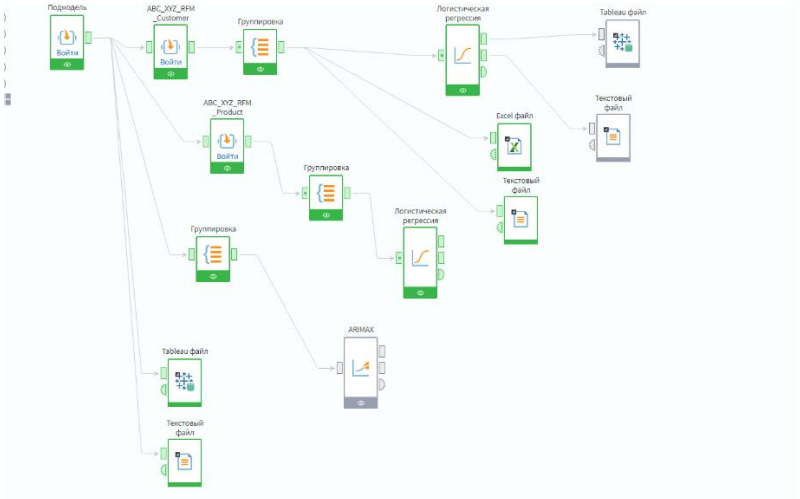


Данные объединены в одну подмодель.

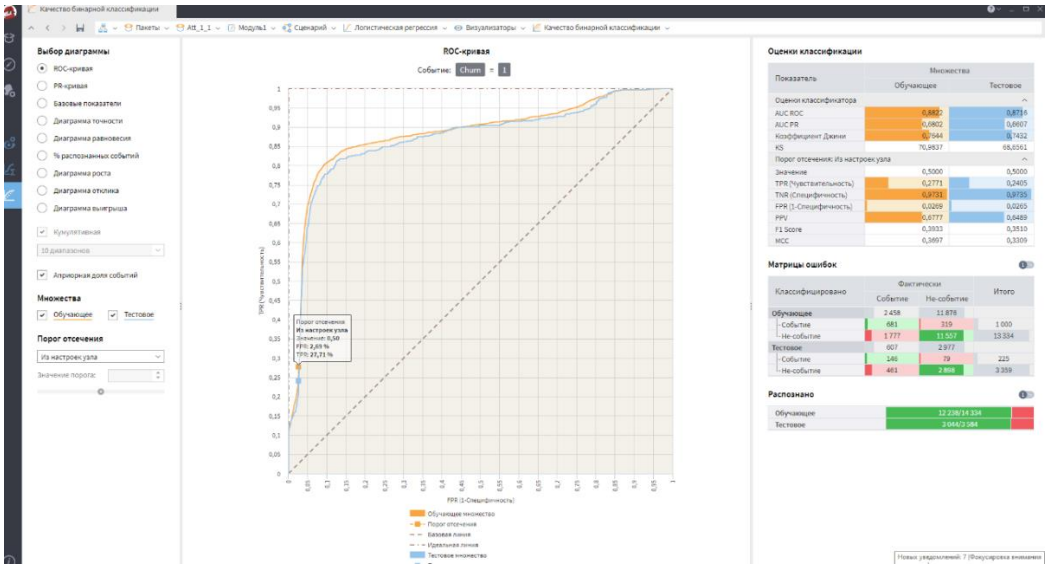
Данный кейс обогащен ABC-XYZ-RFM анализом по клиентам. Для решения задачи бинарной классификации по оттоку клиентов добавлена целевая переменная Churn. Датой события считается 01.04.2016 года. Если максимальная дата покупки раньше даты 01.04.2016, то клиент считается ушедшим.



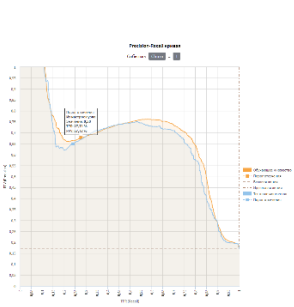
Эти данные также объединены в одну подмодель. Строки сгруппированы по клиенту по максимальной дате заказа. Сгруппированы количество заказов, количество уникальных продуктов в заказах, итоговое количество товаров, суммы и налоги. В начальной таблице клиентов было 18484. Информация по 566 клиентам была удалена. Соответственно осталось 17918 строк для дальнейшей работы и анализа.

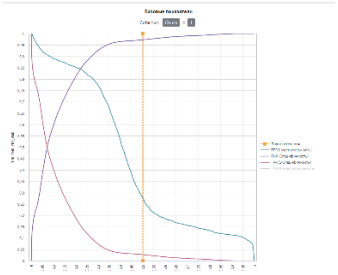


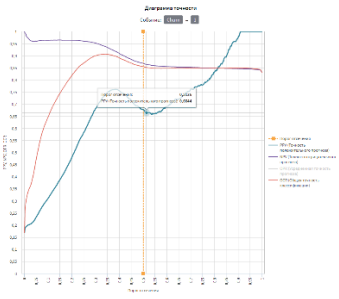
Создана и обучена модель Логистической регрессии. Проведен прогноз.

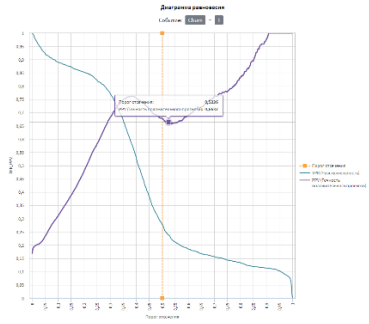


ROC- кривая

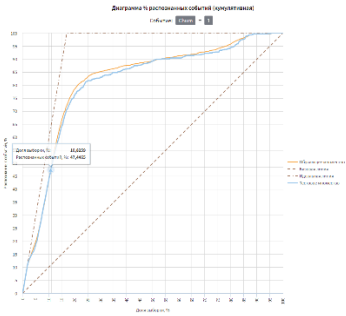


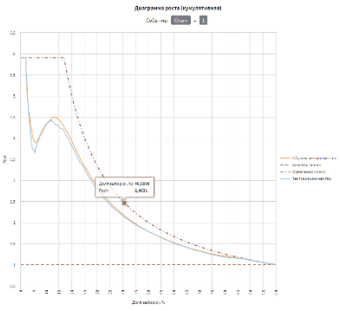


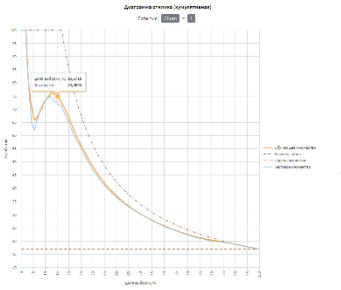


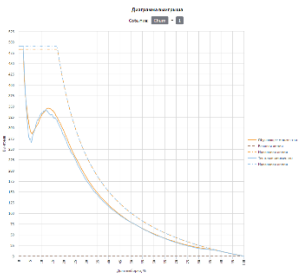


PR- кривая Базовые показатели Диаграмма точности Диаграмма равновесия









% распознанных событий Диаграмма роста Диаграмма отклика Диаграмма выигрыша

Оценки классификации

Показатель	Множества	
	Обучающее	Тестовое
Оценки классификатора		
AUC ROC	0,8822	0,8716
AUC PR	0,6802	0,6607
Коэффициент Джини	0,7644	0,7432
KS	70,9837	68,6561
Порог отсечения: Из настроек узла		
Значение	0,5000	0,5000
TPR (Чувствительность)	0,2771	0,2405
TNR (Специфичность)	0,9731	0,9735
FPR (1-Специфичность)	0,0269	0,0265
PPV	0,6777	0,6489
F1 Score	0,3933	0,3510
MCC	0,3697	0,3309

Матрицы ошибок

Классифицировано	Фактически		Итого
	Событие	Не-событие	
Обучающее	2 458	11 876	
Событие	681	319	1 000
Не-событие	1 777	11 557	13 334
Тестовое	607	2 977	
Событие	146	79	225
Не-событие	461	2 898	3 359

Распознано

Обучающее	12 238/14 334
Тестовое	3 044/3 584

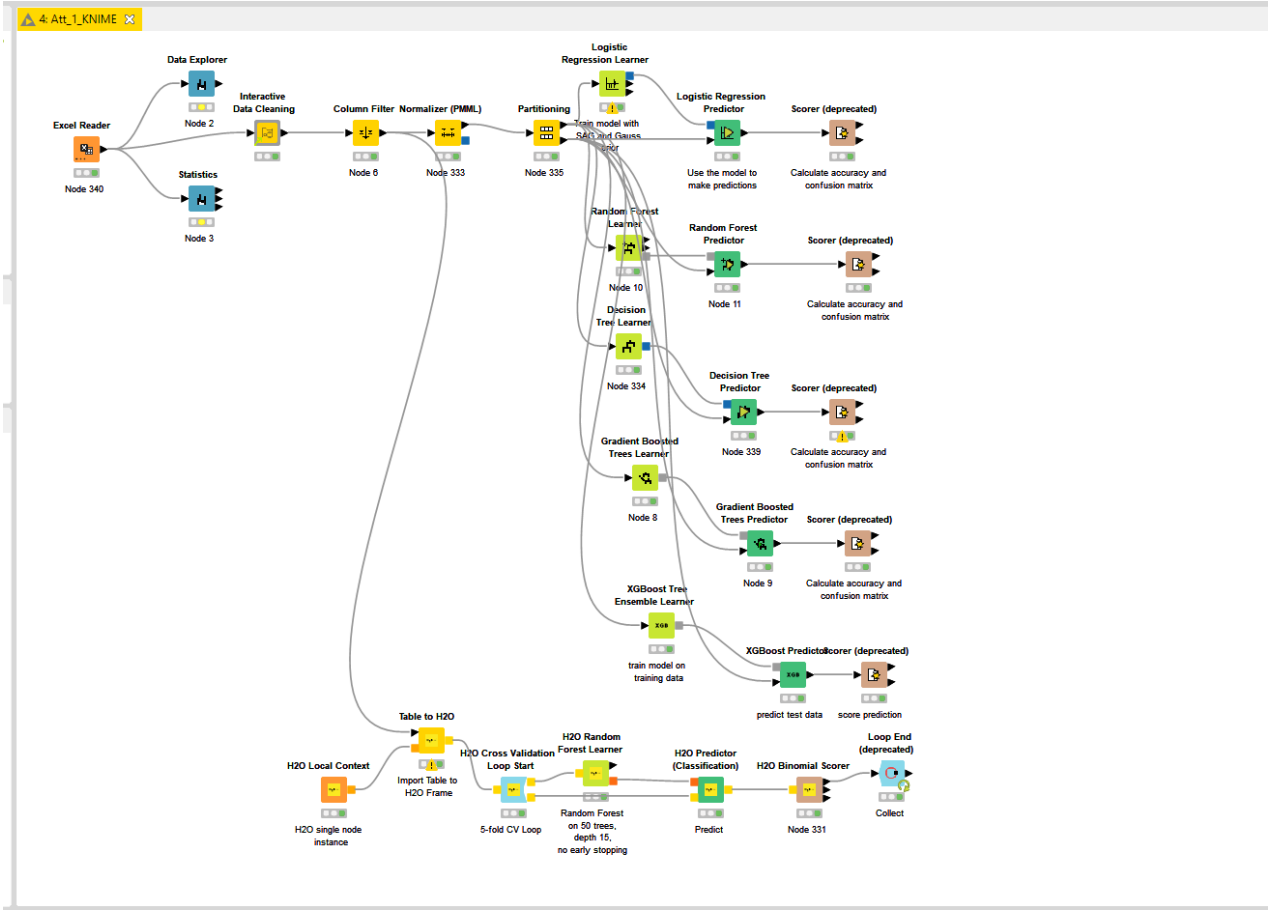
Train	test
AUC-ROC- 0.8822	0.8716
AUC-PR – 0.6802	0.6607
F1-Score – 0.3933	0.3510

Наиболее весовыми признаками в пользу того, что клиент останется по данным Логистической регрессии в Логином являются Количество заказов по клиенту, страны Австралия и Канада, Количество наименований товаров, Наличие детей, Годовой доход.

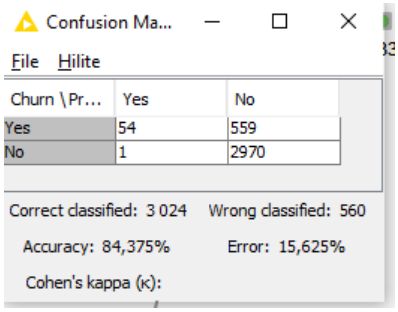
Признаками положительно влияющими на уход клиента являются – дальнее расстояние от работы, страна Германия.

Для создания, обучения и сравнения моделей выгружен файл в формате .csv с данными в количестве 17918 строк обогащенными ABC-XYZ-RFM анализом, с целевой переменной Churn.

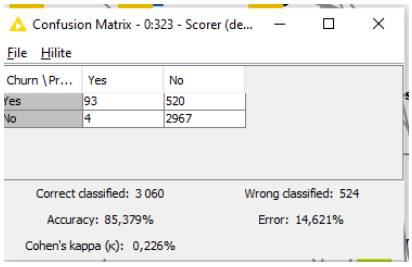
Файл загружен в Knime, созданы модели LogisticRegressionLearner, RandomForestLearner, DecisionTreeLearner, GradientBoostedTreesLearner, XGBoostTreeEnsembleLearner, H2ORandomForestLearner.



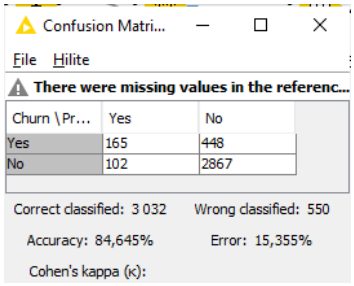
LogisticRegressionLearner



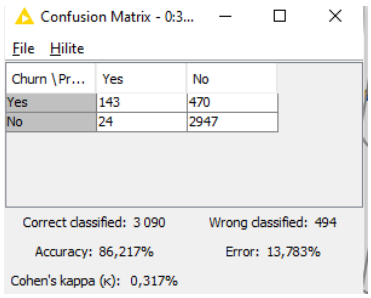
RandomForestLearner



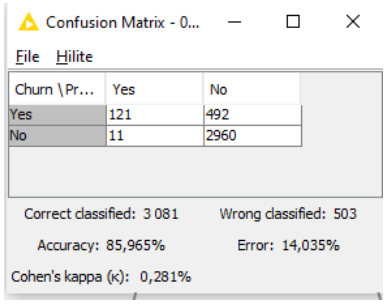
DecisionTreeLearner



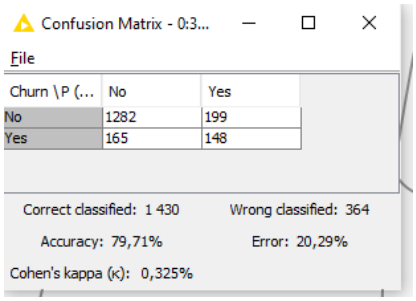
GradientBoostedTreesLearner



XGBoostTreeEnsembleLearner



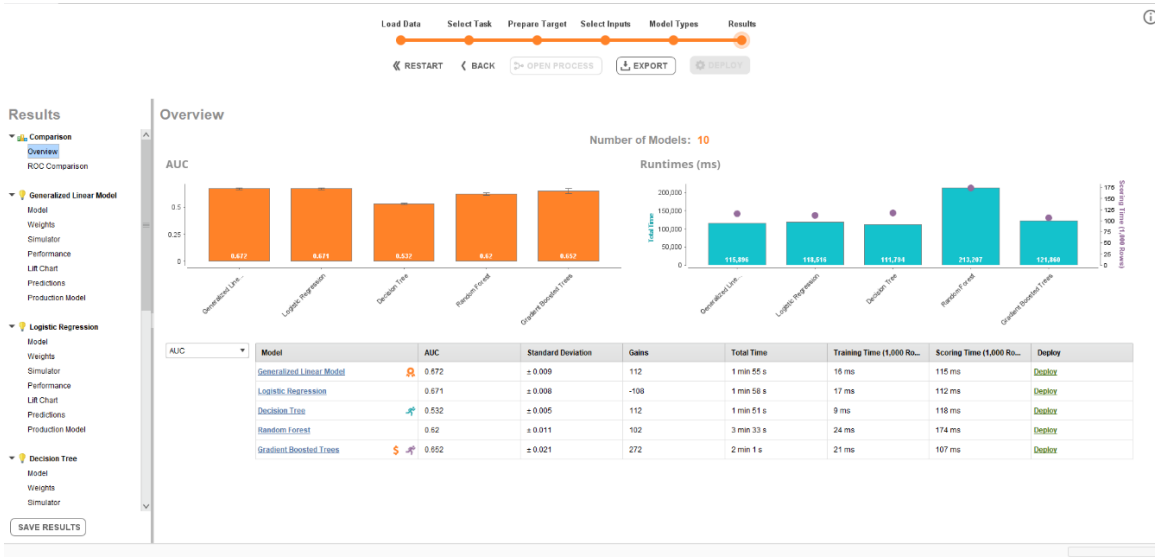
H2ORandomForestLearner

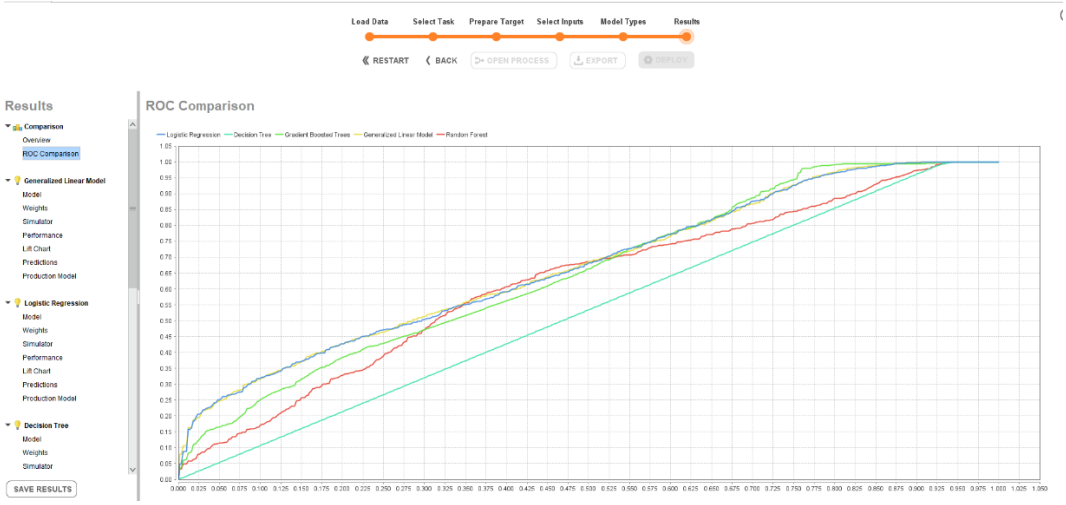


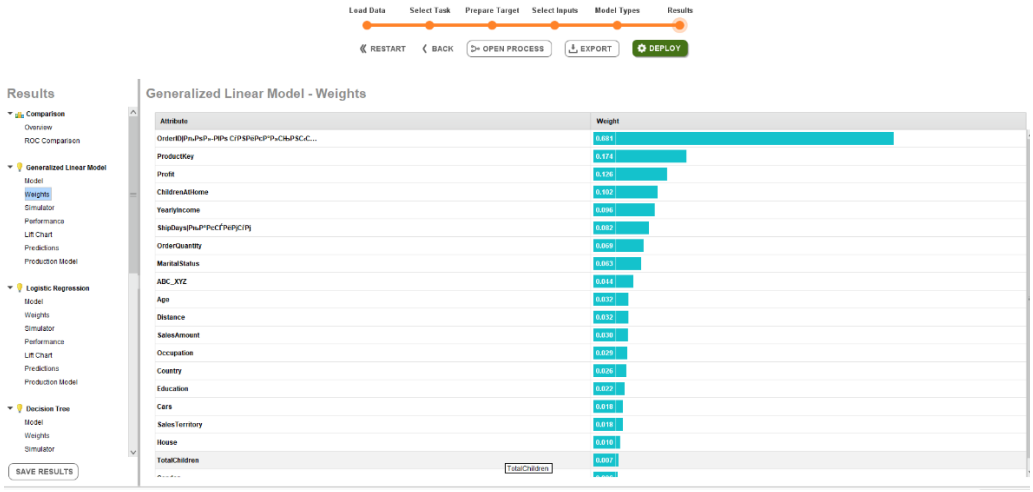
Все модели показывают достаточно хорошие схожие высокие метрики, но лучшие показатели все же у GradientBoostedTreesLearner - Accuracy – 86.217 и самый низкий показатель по матрице ошибок 13,783%.

Этот же датасет загружен в RapidMiner

Обучены модели Generalized Linear Model, Logistic Regression, DecisionTree, GradientBoostedTrees







Модели, обученные в RapidMiner, как и в Knime, имеют схожие метрики, но лучшие показатели у Generalized Linear Model - AUC-score – 0.672. Самые важные признаки у всех моделей определены идентично – это количество заказов, разнообразие заказываемых товаров, отсутствие детей, годовой доход. Чем выше коэффициенты у данных показателей, тем вероятнее, что клиент останется.

Самые лучшие показатели у модели, обученной при помощи библиотеки scikit-learn-GradientBoostingClassifier – ROC_AUC Score – 0.8727. У моделей, обученных при помощи библиотек dabl – LogisticRegression – ROC_AUC Score 0.689. Модели библиотек dabl показывают близкие результаты к результатам моделей RapidMiner.

Результаты по моделям победителям из разных платформ по ROC_AUC Score.

Loginorm –	LogisticRegression	-0.87
Knime –	GradientBoostedTreesLearner	-0.86 (Accuracy)
RapidMiner –	Generalized Linear Model	-0.67
Scikit-learn –	GradientBoostingClassifier	-0.76
Dabl -	LogisticRegression	-0.689

Основываясь на исследованиях и анализе автоматических визуализаций и визуализаций в POWER BI, можно сказать, что явных зависимостей и предрасположений к уходу клиентов нет. Ни возраст, ни пол, ни доход, ни образование и род занятий особого значения не имеют.

Есть явная зависимость от количества товаров в заказе, от разнообразия покупаемых товаров, от расстояния между домом и работой, а также явно видна зависимость от наличия детей.

Магазин, расширяя и увеличивая ассортимент товаров и категорий, привлекает новых клиентов и способен удержать имеющихся. Стоит уделить внимание молодым клиентам. Не стоит оставлять без внимания тот факт, что среди покупателей нет клиентов моложе 40 лет. Возможно стоит заняться расширением линейки товаров с ориентацией на детскую аудиторию, тем самым будут привлечены и покупатели, имеющие детей. Так же стоит обратить внимание на продажи своих товарам в соседние штаты и города. В США, где самые большие объемы продаж, охвачено только три штата. Также и в Австралии.