

AI 21 – Uczenie maszynowe - Lotnisko i reklama (13)

Orange Data Mining (ODM) to darmowy, otwartoźródłowy program służący do wizualnej analizy danych, uczenia maszynowego oraz eksploracji danych. Program ten wyróżnia się interfejsem graficznym, który umożliwia budowanie zaawansowanych procesów analitycznych za pomocą przeciągania i łączenia klocków, zwanych widgetami, bez konieczności pisania kodu.

Podstawowe widgety:

Data - wczytywania i filtrowania danych,

Transform - przekształcania tabel z danymi,

Visualize - tworzenia interaktywnych wykresów,

Model - maszynowego uczenia na podstawie danych,

Evaluate - testowania jakości wyników, analiza tekstu, przewidywanie, itp.

Pamiętaj o tym, by zrzut ekranu DOKUMENTOWAŁ Twoją pracę

ZADANIE

Przeanalizuj liczbę pasażerów na polskich lotniskach od 2020 roku do dzisiaj.

Przygotuj tabelę (asystent AI), na podstawie której dokonasz analizy w programie ODM.

1. Analiza jakościowa

- Uruchom dowolnego asystenta AI

np.  Tryb AI  Gemini  Microsoft 365 Copilot  ChatGPT

- Zadaj asystentowi AI pytanie:

Przeanalizuj
liczbę pasażerów na polskich lotniskach od 2000 roku do czasów obecnych.
Wypisz 4 najważniejsze wnioski z tych obserwacji w krótkich zdaniach

- Wklej do ramki zrzut ekranu wniosków z analizy której dokonała AI



2. Dane

```
Data,Liczba pasazerow
2000-01,335000
2000-02,321000
2000-03,365000
2000-04,410000
(1) 2000-05,480000
```

- Za pomocą asystenta AI:

przygotuj tabelę z ruchem pasażerskim na polskich lotniskach od 2000 roku.
Tabela ma składać się z dwóch kolumn:
- pierwsza kolumna: data w postaci rok-miesiąc
- druga kolumna: liczba pasażerów
Dane mają być w formacie kodu CSV do skopiowania

- Skopiuj wszystkie dane 
- Wklej do notatnika

Nazwa pliku: libront.csv

Zapisz jako typ: All types (*.*)

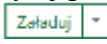
- Zapisz notatnik w swoim folderze z nazwą **nazwisko.csv**
- *UWAGA - wpisz oczywiście swoje nazwisko*
- Wklej do ramki zrzut ekranu z tabelą
- Zamknij notatnik



Masz problem z asystentem AI - pobierz gotowy plik: http://hosting18.vpx197.iat.pl/pliki/ruch_pasazerski.csv

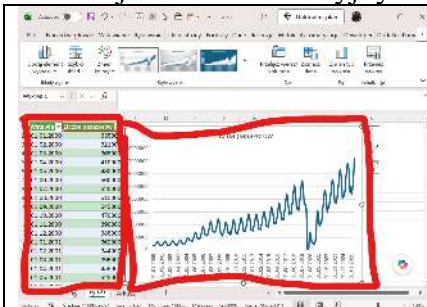
3. Wykres na arkuszu

- Otwórz arkusz kalkulacyjny (Excel)
- Wybierz z menu: **Dane - Pobierz dane - Z pliku - Tekstowego/CSV**
- Odszukaj i wczytaj plik **nazwisko.csv**

Wciśnij przycisk 

- Zaznacz obie kolumny z tabelą

- Wybierz z menu: **Wstawianie - Wykres liniowy**  - **Liniowy 2D** 
- Wklej do ramki zrzut ekranu z tabelą i wykresem
- Zamknij arkusz kalkulacyjny



4. Wczytaj dane

Cell delimiter: Comma
Quote character:
Number separators: Grouping
Column type:

	1	2
1	Data	Herba pasażerów
2	2000-01	335000
3	2000-02	321000
4	2000-03	365000

(1)

(2)

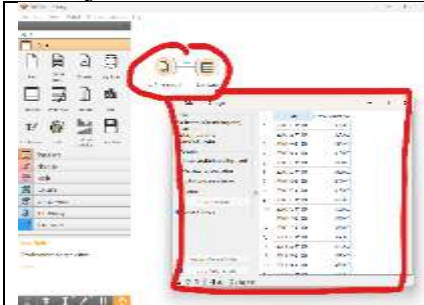


- Uruchom program **Orange data Mining (ODM)** 

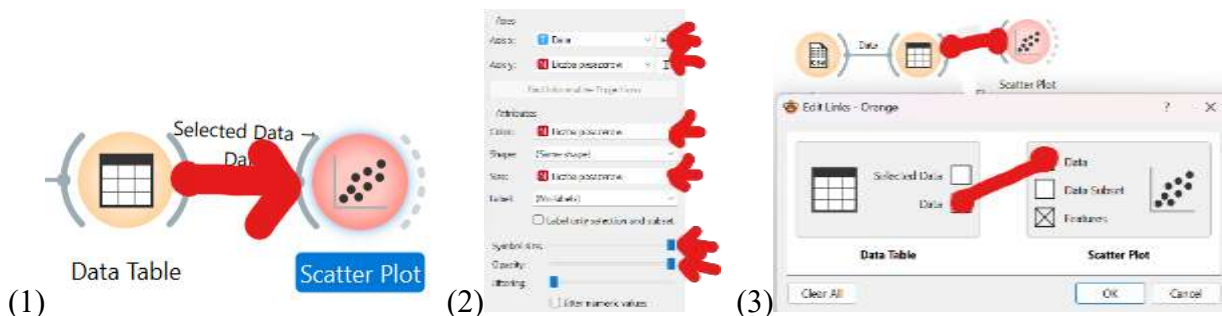
- Z modułu **Data**  wstaw widget **CSV File Import** 

- Otwórz widget
Kliknij podwójnie

- Wybierz przycisk 
- (1) Odszukaj na dysku i otwórz plik **csv** z danymi o lotach
UWAGA - jeśli nie widać tabeli z osobnymi kolumnami, spróbuj zmienić przecinek Comma na średnik Semicolon
- Zamknij okienko **Import Options** przyciskiem **OK**
- Zamknij okienko z widgetu **CSV File Import** przyciskiem **X**
- Z modułu **Data**  **Data** wstaw widget **Data Table** 
- (2) Połącz widgety **CSV File Import** z **Data Table**
Złap kropkę jednego widgetu i przeciągnij do kropki drugiego widgetu
- Otwórz widget **Data Table**
Powinna pojawić się tabela z danymi o lotach
- Wklej do ramki zrzut ekranu: widać połączone widgety i tabelę



5. Wykres punktowy



- Z modułu **Visualize**  **Visualize** wstaw widget **Scatter Plot** 
- (1) Połącz **Data Table** z **Scatter Plot**
- Otwórz **Scatter Plot**
- (2) Ustaw parametry wykresu **Scatter Plot**
Oś X: Data
Oś Y: Liczba pasażerów
Color i Size: Liczba pasażerów
im więcej pasażerów, tym większa i jaśniejsza kropka
- (3) Kliknij podwójnie w linię łączącą widgety **Data Table** i **Scatter Plot**
- Ustaw połączenie: **Data - Data** i wciśnij przycisk **OK**
Jeśli jest ustawione poprawnie - nic nie zmieniaj
Połączenie Selectet Data - Datapozwala pokazywać na wykresie tylko zaznaczone w tabeli wiersze
- Wklej do ramki zrzut ekranu z wykresem



6. Szereg czasowy

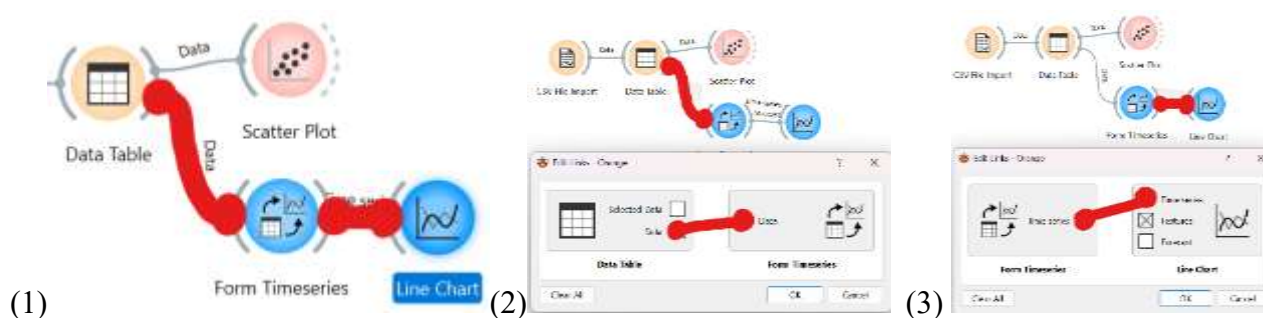
Moduł Timeseries w programie Orange Data Mining jest opcjonalnym dodatkiem (add-on), który zawiera zestaw wyspecjalizowanych widżetów (klocków) dedykowanych do pobierania, przekształcania, wizualizacji oraz modelowania szeregów czasowych. Moduł ten dzieli się na cztery główne kategorie narzędzi:

Pobieranie i transformacja danych (Data & Preprocessing): zamienia zwykłe tabele na szeregi czasowe

Wizualizacja (Visualization): wykresy

Modelowanie i Prognozowanie (Modeling) ARIMA Model: przewidywanie przyszłości

Ocena modeli (Evaluation) Model Evaluation: testuje dopasowanie modeli i mierzy błędy prognoz przy zachowaniu chronologii danych (zapobiega wyciekowi danych z przyszłości do przeszłości).

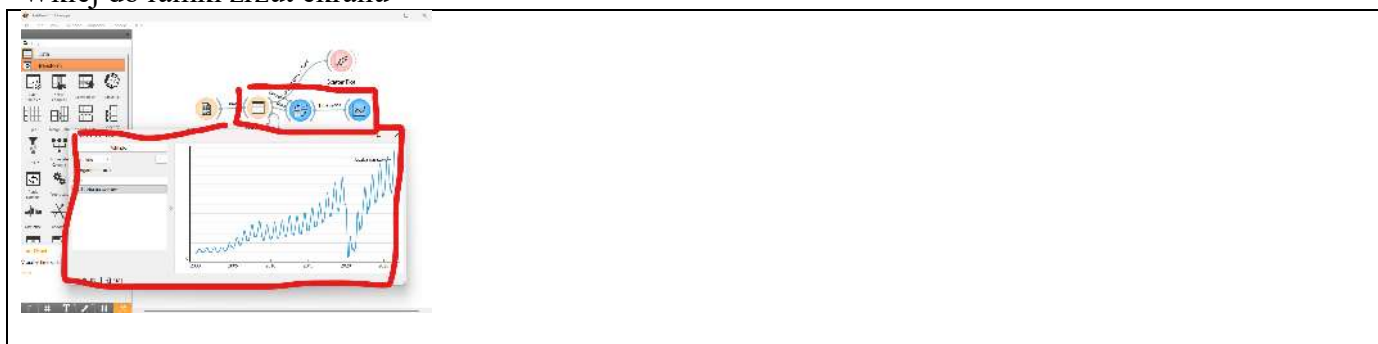


- Z modułu **Time Series** wstaw dwa widżety:

LineChart i **Form Timeseries**

UWAGA - jeśli moduł nie jest widoczny, wybierz: Options - Add-ons - ☒ Timeseries

- (1) Połącz widżety
 - (2) Kliknij podwójnie w linię połączenia **Data Table - Form Timeseries**
- ustaw **Data - Data**
 - (3) Kliknij podwójnie w linię połączenia **Form Timeseries - Line Chart**
- ustaw **Time Series - Time Series**
- UWAGA - połączenie między widgetami ustawiają się automatycznie i prawie zawsze poprawnie*
- Otwórz widżet **Line Chart**
Wykres liniowy liczby pasażerów
 - Wklej do ramki zrzut ekranu



7. Prognozowanie

Prognozowanie szeregów czasowych (Time Series Forecasting) w Orange Data Mining służy do przewidywania przyszłych wartości (np. liczby pasażerów, sprzedaży) na podstawie chronologicznych danych historycznych. Program analizuje trendy i sezonowość z przeszłości, aby wygenerować prognozę na określony czas w przód.

Kluczowe są dwa widżety:

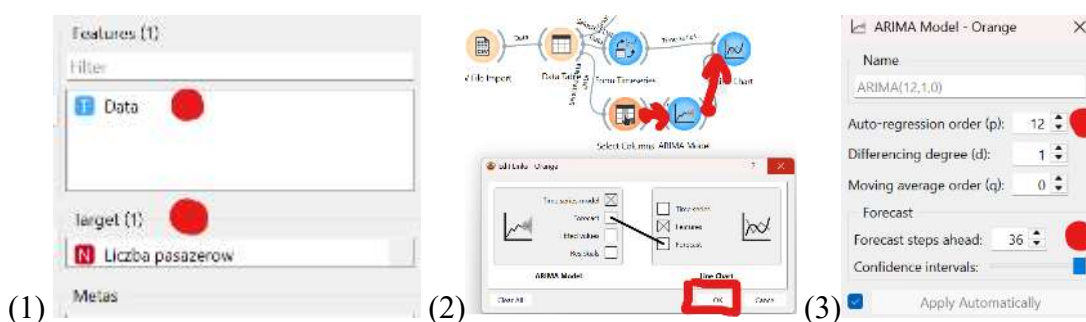
Form Timeseries – przekształcenie tabeli w chronologiczny szereg czasowy (wskazanie osi czasu)

ARIMA - algorytm prognozujący, który opiera swoje prognozy wyłącznie na matematycznych zależnościach

W widżecie **Line Chart** programu Orange Data Mining, po podłączeniu modelu ARIMA, niebieska linia oraz otaczające ją jasnoniebieskie pole oznaczają prognozę (szacunek punktowy) oraz jej przedział ufności (margines błędu).

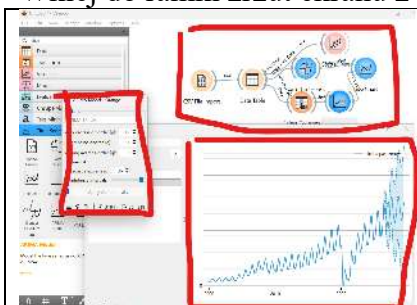
Niebieska linia (Forecast / Prognoza punktowa): Jest to najbardziej prawdopodobna ścieżka, jaką według modelu ARIMA przyjmie badana zmienna w przyszłości. Algorytm bierze pod uwagę historyczne trendy, cykle (autoregresję) oraz ostatnie wstrząsy losowe i wylicza jedną, konkretną wartość dla każdego kolejnego kroku w przód (np. na kolejne 10 dni).²

Jasnoniebieskie pole (Confidence Interval / Przedział ufności): Jest to obszar niepewności prognozy. Pokazuje zakres, w którym rzeczywiste (przyszłe) wartości powinny się znaleźć z określonym prawdopodobieństwem (standardowo wynosi ono w Orange 95%). Im dalej w przyszłość sięga prognoza, tym pole staje się szersze (przybiera kształt lejka). To całkowicie naturalne – im dalsza przyszłość, tym trudniej ją przewidzieć i tym większe ryzyko błędu. Pierwsze 2-3 kroki prognozy są zazwyczaj dość wąskie (bardziej pewne), a dalsze kroki stają się bardzo szerokie (mniej pewne).



- Z modułu **Transform**  wstaw widżet **Select Columns** 
- Połącz widżet **Data Table** z **Select Columns**
- połączenie: **Data - Data**
- (1) Otwórz **Select Columns** i ustaw:
Features: Data - dane wejściowe wykorzystywane do obliczania, przewidywania, itp.
Target: Liczba pasażerów - zmienne wyjściowe, docelowe, obliczone, przewidziane, itp.
Metas - nie biorą udziału w obliczeniach - są pomijane

- Z modułu **Time Series** wstaw widżet **ARIMA Model** 
- (2) Połącz go z widżetem **Select Columns** z jednej strony, a z drugiej z **Line Chart**
- połączenie: **Forecast - Forecast**
- Otwórz widżet **ARIMA**
- (3) Ustaw parametry ARIMA, jak na obrazku
Auto-regression order - 12
automatyczne dopasowanie do liczby miesięcy
Forecast step ahead - 36
przewidywanie na 36 miesięcy do przodu
- Otwórz widżet **Line Chart**
niebieska linia oraz otaczające ją jasnoniebieskie pole oznaczają prognozę (szacunek punktowy) oraz jej przedział ufności (margines błędu)
- Wklej do ramki zrzut ekranu z widżetami **LineChart** i **ARIMA**



8. Przedział ufności

- Zapytaj asystenta AI

co oznacza niebieskie pole i niebieska przerywana linia na wykresie Line Chart w programie Arange Data Mining

- Wklej do ramki zrzut ekranu z odpowiedzi, której udzielił asystent AI

- Na wykresie **Line Chart** kliknij w pole **Time Series**

W lewym dolnym rogu wykresu

- Odszukaj wartości, które przewiduje model ARIMA dokładnie za rok (za 12 miesięcy)

UWAGA: sekcja Forecast

- Wpisz do ramki:

- wartość prognozy punktowej

- szerokość przedziału ufności

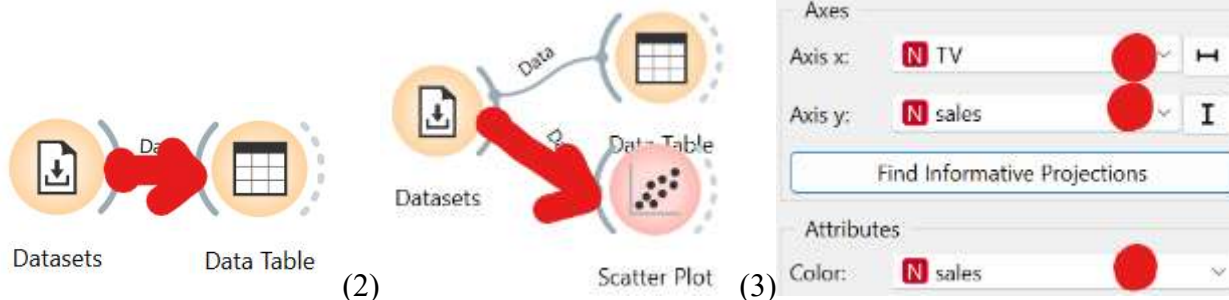
Różnica maksymalnej i minimalnej wartości niebieskiego pola

	
wartość prognozy punktowej po roku	-
szerokość przedziału ufności	-

Reklama

Tabela Advertising (Reklama) w programie Orange Data Mining to bardzo popularny, klasyczny zbiór danych służący do nauki Regresji Liniowej oraz optymalizacji budżetów marketingowych. Pochodzi z cenionego podręcznika akademickiego *An Introduction to Statistical Learning (ISLR)*. Głównym zadaniem algorytmu jest przewidzenie wielkości sprzedaży produktu (zmienna Sales) na podstawie nakładów finansowych przeznaczonych na reklamę w trzech różnych kanałach medialnych: TV, radio, gazety.

9. Dane



- W programie Orange Data Mining przygotuj nowy dokument
File - New

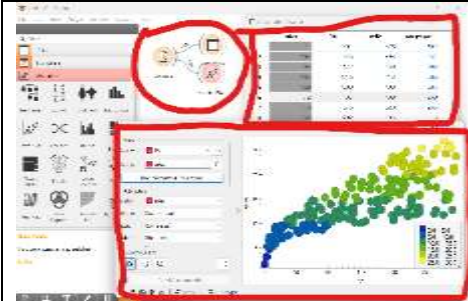
- Z modułu **Data** wstaw widget **Datasets**

- Otwórz widget i odszukaj tabelę o nazwie **Advertising**

Advertising	15.5 KB	200	4	N numeric	economy
-------------	---------	-----	---	-----------	---------

- Kliknij podwójnie w tabelę (aby ją załadować)

- Zamknij okienko widgetu
Tabela zawiera 200 wierszy i 4 kolumny (zmienne)
Kolumny **TV**, **Radio** oraz **Newspaper** (budżety reklamowe) mają przypisaną rolę *feature* (zmienne objaśniające)
Kolumna **Sales** (wartość sprzedaży) ma przypisaną rolę *target* (zmienna objaśniana / cel)
- (1) Z modułu **Data** wstaw widget **Data Table**
- Połącz go z **Datasets**
typ połączenia Data - Data
- (2) Z modułu **Visualize** wstaw widget **Scatter Plot**
- Połącz go z **Data Table**
typ połączenia Data - Data
- Otwórz **Data Table**
- (3) Otwórz **Scatter Plot** i ustaw osie i atrybuty
Kwoty przeznaczone na reklamę TV (oś X) i wartość sprzedaży (oś Y)
- Wklej do ramki rzut ekranu: widać połączenia, tabelę i wykres



10. Szybka analiza

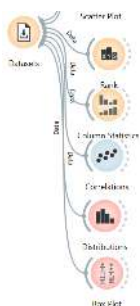
Algorytm **Rank** przegląda wszystkie dostępne kolumny (TV, Radio, Newspaper) i układa je w kolejności od tej, która ma największy wpływ na Twój cel (Sales), do tej, która jest całkowicie bezużyteczna.

Widget **Column Statistic** pokazuje histogram (dane rozkładają się równomiernie, czy nie zawierają błędów), a także minimum, maksimum, średnią i najczęstszą wartość.

Widget **Correlations** sprawdza, czy gdy jedna wartość w tabeli rośnie (np. budżet na TV), to druga wartość również się zmienia (np. sprzedaż rośnie lub spada). W naszym przypadku największy wpływ na wartość sprzedaży ma TV.

Widget **Distributions** sprawdza, jak często poszczególne wartości pojawiają się w danych (histogram). Równomierny rozkład nie wyróżnia żadnej konkretnej wartości, kształt dzwonowaty pozwala wybrać tę, która przynosi największy zysk, rozstrzelone wartości wskazywać mogą na błędne dane.

Widget **Box Plot** - podstawowe dane statystyczne w ujęciu graficznym. Linia środkowa (mediana) to idealny środek danych, niebieskie pudełko zawiera 50% wszystkich wyników (najczęstsze), Wąsy - minimalne i maksymalne wartości, pojedyncze kropki - wartości odstające, anomalie (błędne)



(1)

- (1) Z modułu **Data** wstaw widget **Rank**  i połącz go z **Datasets**
- Z modułu **Data** wstaw widget **Column Statistic**  i połącz go z **Datasets**
- Z modułu **Unsupervised** wstaw widget **Correlations**  i połącz go z **Datasets**

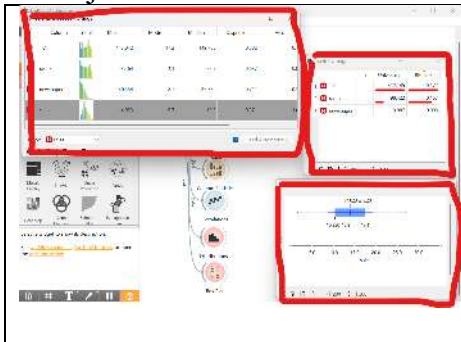
- Z modułu **Visualize** wstaw widget **Distributions** i połącz go z **Datasets**



- Z modułu **Visualize** wstaw widget **Box Plot** i połącz go z **Datasets**



- Otwórz wszystkie widżety do szybkiej analizy i rozmieść je na ekranie
Możesz pozamykać panele kontrolne z lewej strony każdego widżetu
- Wklej do ramki zrzut ekranu



11. Modele uczenia maszynowego AI

Widżet Test & Score (Testowanie i Ocena) to najważniejsze narzędzie w programie Orange służące do sprawdzania, jak dokładne i wiarygodne są Twoje modele sztucznej inteligencji (takie jak regresja liniowa czy las lasowy). Działa on jak surowy egzaminator: bierze model, daje mu zadanie do rozwiązywania i wystawia ocenę końcową.

Kluczowe oceny:

Cross-Validation (Walidacja krzyżowa): to najbardziej rzetelna i zalecana metoda oceny modelu. Program dzieli cały zbiór danych na określoną liczbę równych części, np. 10 (tzw. 10-fold Cross-Validation). Następnie wykonuje 10 rund testowych. W każdej rundzie 9 części służy do uczenia modelu, a pozostała 1 część do jego przetestowania. Proces powtarza się, aż każda część była dokładnie raz zbiorem testowym.

Random sampling (Losowe próbkowanie): to metoda sprawdzania skuteczności modeli. Program działa jak losowanie loterii: dzieli dane na jedną dużą część do nauki i jedną mniejszą do egzaminu, a cały proces może powtórzyć wielokrotnie. Szukamy modelu testowania danych, który ma największe R² a najmniejsze RMSE

R² (R-kwadrat): Pokazuje w procentach (0.00 - 1.00), jak dobrze wydatki na reklamy tłumaczą sprzedaż. Wynik 0.89 oznacza, że model wyjaśnia aż 89% rynku.

RMSE / MAE: To średni błąd prognozy. Jeśli RMSE wynosi np. 1.5, oznacza to, że Twój model prognozując sprzedaż, myli się średnio o 1.5 tysiąca sztuk towaru.

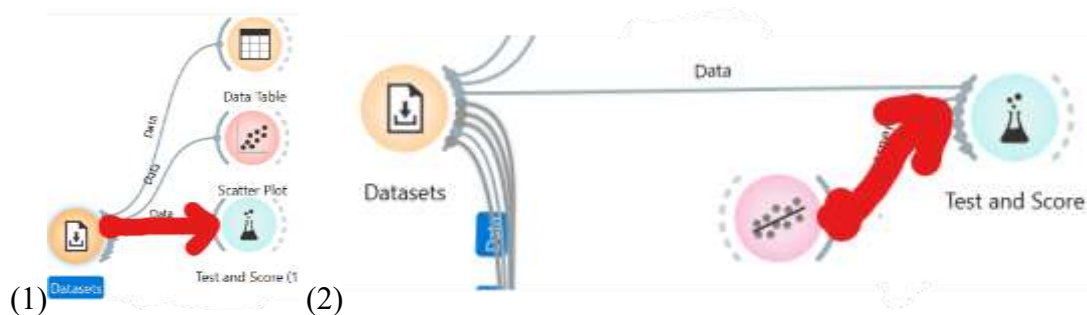
Linear Regression (Regresja liniowa) służy do znajdowania prostego wzoru matematycznego, który łączy Twoje wydatki na reklamy z końcową sprzedażą. Działa jak linijka, która próbuje narysować idealnie prostą, uniesioną linię przechodzącą przez środek chmury kropek na wykresie.

Tree (Drzewo decyzyjne) służy do prognozowania wyników na podstawie prostych, logicznych reguł postępowania. W przeciwieństwie do skomplikowanego Lasu Losowego (Random Forest), to narzędzie buduje tylko jedno, główne drzewo, które możesz w całości obejrzeć na ekranie komputera. Działa ono jak algorytm zadający pytania: sprawdza tabelę Advertising, szuka najważniejszego progu finansowego i dzieli dane na gałęzie (np. „Czy budżet na TV jest większy niż 100 tys.?” → Jeśli TAK, idź w prawo; jeśli NIE, idź w lewo).

Random Forest (Las losowy): w przeciwieństwie do regresji liniowej, która szuka jednej prostej linii, las działa jak komitet ekspertów. Tworzy w tle kilkadziesiąt lub kilkaset niezależnych drzew decyzyjnych. Każde z tych drzew analizuje tabelę Advertising pod innym kątem, a końcowa prognoza sprzedaży jest średnią ze wszystkich ich opinii.

Gradient Boosting (Wzmocnienie gradientowe) to jeden z najpotężniejszych i najnowocześniejszych modeli uczenia maszynowego. W przeciwieństwie do modelu Random Forest (gdzie wszystkie drzewa powstają niezależnie obok siebie), Gradient Boosting buduje drzewa sekwencyjnie – jedno po drugim. Każde kolejne drzewo ma jedno zadanie: naprawić błędy popełnione przez swojego poprzednika.

SVM (Support Vector Machines, czyli Maszyny Wektorów Nośnych) to zaawansowany algorytm, który szuka optymalnego „tunelu” wokół linii trendu. Stara się zmieścić jak najwięcej punktów danych wewnątrz tego tunelu, ignorując drobne błędy, a skupiając się na punktach leżących na jego brzegu (tzw. wektorach nośnych).



- (1) Z modułu **Evaluate** wstaw widget **Test and Score** 
i połącz go z **Datasets**
Polączenie Data - Data

- (2) Z modułu **Model** wstaw widget **Linear Regresion** 
i połącz go z **Test and Score**
Polączenie Learner - Learner

- Z modułu **Model** wstaw widget **Tree** 
i połącz go z **Test and Score**
Polączenie Learner - Learner

- Z modułu **Model** wstaw widget **Random Forrest** 
i połącz go z **Test and Score**
Polączenie Learner - Learner

- Z modułu **Model** wstaw widget **Gradient Boosting** 
i połącz go z **Test and Score**
Polączenie Learner - Learner

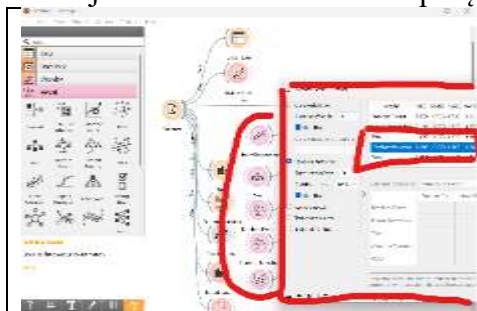
- Z modułu **Model** wstaw widget **SVM** 
i połącz go z **Test and Score**
Polączenie Learner - Learner

- Otwórz Widget **Test and Score**

- Ustaw **Random sampling**  **Random sampling**

- Zaznacz model, który jest najbardziej wiarygodny dla naszych danych
- ma najwyższe wartości w tabeli

- Wklej do ramki zrzut ekranu z połączeniami i oceną modeli

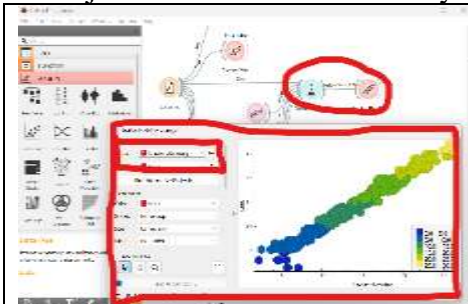


12. Ocena wizualna

Zamiast porównywać liczby można zwizualizować błędy i ocenić jakość prognoz za pomocą wykresu

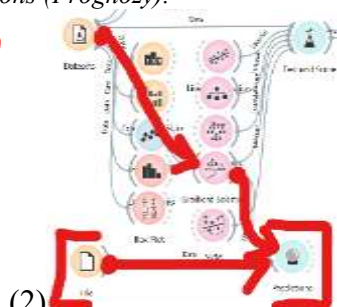
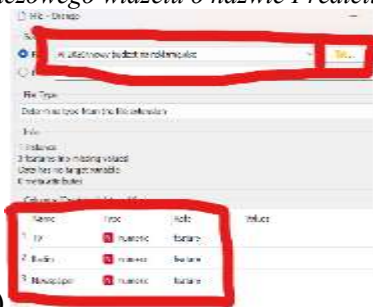


- (1) Z modułu **Visualize** wstaw widget **Scatter Plot** i połącz z **Test and Score**
Połączenie: Predictions - Data
- Otwórz widget **Scatter Plot** i ustaw osie
X: Gradient Bosting
Y - Sales
- Zmień model na osi X - sprawdź który jest najmniej „rozstrzelony”
Im mniej rozstrzelone wartości, tym model oceniania lepszy
- Wklej do ramki zrzut ekranu z wykresem dla modelu **Gradient Bosting**



13. Prognozowanie sprzedaży - Nowe dane

Prognozowanie sprzedaży (ang. Sales Forecasting) przy użyciu modeli uczenia maszynowego w Orange to główny cel biznesowy pracy z tabelą. Polega ono na wykorzystaniu nauczonych modeli (np. Linear Regression czy Random Forest) do precyzyjnego obliczenia, ile towaru firma sprzedaje w przyszłości po wpisaniu do budżetu konkretnych kwot marketingowych. Aby uruchomić taką automatyczną symulację, musisz przygotować plik z danymi sprzedaży (najlepiej w Excelu) oraz użyć kluczowego widżetu o nazwie Predictions (Prognozy).



- (1) Uruchom arkusz kalkulacyjny (Excel)

	A	B	C
1	TV	Radio	Newspaper
2	140	35	5

- Wpisz dane
UWAGA - nazwy kolumn muszą być identyczne jak wzorcową tabelą
- Zapisz plik w swoim folderze z dowolną nazwą
- Z modułu **Data** wstaw widget **File**
- (1) Otwórz widget **File**
i wczytaj ... zapisany przed chwilą arkusz
W oknie powinny pojawić się nazwy kolumn identyczne jak w oryginalnej tabeli
- (2) Z modułu **Evaluate** wstaw widget **Predictions**
i połącz **File** z **Predictions**
Połączenie Data - Data

- Połącz **Gradient Boosting** z **Predictions**
Połączenie Model - Predictors
- Połącz **Gradient Boosting** z **Dataset**
Połączenie Data - Data
- Otwórz widжет **Predictions**, aby zobaczyć prognozę sprzedaży
W nowej kolumnie pojawi się prognozowana (na podstawie wybranego modelu) wartość sprzedaży
- Wklej do ramki zrzut ekranu z prognozą sprzedaży

