

RAPORT PROJEKTU ZESPOŁOWEGO Z EKSPŁORACJI DANYCH

Autorzy:

Brygida Juszczuk,

Edyta Margol,

Aleksandra Wójcik

Niniejszy raport przedstawia analizę danych przeprowadzoną w ramach projektu, którego celem było zbadanie zastosowania różnych metod analizy danych w kontekście grupowania i prognozowania. Realizacja projektu została podzielona na trzy zadania, każde z nich wykorzystujące inne techniki analizy i modelowania danych. Poniżej przedstawiono szczegóły dotyczące każdego z zadań oraz przyjęte podejście analityczne.

Zadanie 1.

Celem pierwszego zadania było przeprowadzenie grupowania danych z wykorzystaniem trzech różnych technik:

- Drzewa decyzyjnego - jego działanie polega na rekurencyjnym podziale danych na mniejsze, bardziej jednorodne podzbiory, tworząc strukturę przypominającą drzewo. Proces budowy drzewa rozpoczyna się od całego zbioru danych i polega na wyborze najlepszych cech oraz wartości progowych do podziału, co maksymalizuje jednorodność podzbiorów. Węzły wewnętrzne drzewa reprezentują decyzje na podstawie wybranych cech, natomiast liście reprezentują końcowe przewidywania modelu. Algorytm dzieli dane aż do osiągnięcia określonej głębokości drzewa, minimalnej liczby próbek w liściu lub gdy dalsze podziały nie przynoszą znaczącej poprawy jednorodności,
- Gradient boostingu - polega na budowaniu sekwencji drzew decyzyjnych w sposób iteracyjny. Każde kolejne drzewo w serii poprawia błędy poprzednich, ucząc się na podstawie różnic (residuum) między przewidywaniami a rzeczywistymi wartościami,
- Lasu losowego - polega na tworzeniu wielu drzew działających jako zespół. Każde z tych drzew jest trenowane na losowo wybranym podzbiorze danych oraz losowym podzbiorze cech, co wprowadza różnorodność w drzewach. Przewidywania poszczególnych drzew są następnie łączone – w przypadku klasyfikacji poprzez wybranie najczęściej powtarzającej się wartości, a w przypadku regresji poprzez uśrednianie wyników. Las losowy jest bardziej odporny na szum w danych i zazwyczaj oferuje lepszą wydajność w porównaniu do pojedynczego drzewa decyzyjnego,

oraz dla różnych odsetków zbioru uczącego:

- 70%,
- 80%,
- 90%.

Dodatkowo, przy podziale na zbiór uczący i testowy zastosowano dwa podejścia:

1. **Podział wieloetapowy:**

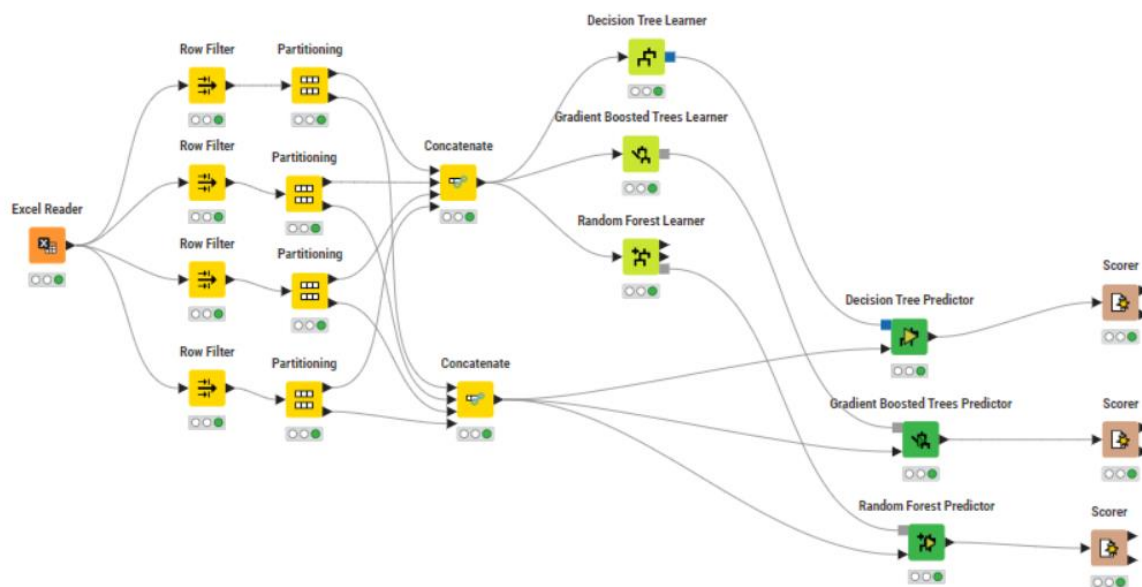
- Najpierw zbiór został podzielony na grupy odpowiadające poszczególnym klasom występującym w zbiorze, gdzie każda grupa reprezentuje jedną klasę.
- Następnie każdej z tych grup wybrano odpowiednio zbiór uczący i testowy, które połączono w jeden zbiór uczący i testowy.

2. **Podział bezpośredni:**

- Cały zbiór danych został podzielony od razu na zbiór testowy i uczący.

Podział wieloetapowy:

Podejście z wykorzystaniem podziału wieloetapowego zostało przedstawione na poniższym rysunku.



Budowa tych modeli opierała się na następujących etapach:

1. Wczytanie danych za pomocą Excel Reader.
2. Wyodrębnienie z danych czterech grup za pomocą Row Filter (ponieważ w zbiorze danych istnieje taka zmienna jak grupa i ona przyjmuje cztery wartości).

3. Z każdej grupy wybranie zbioru uczącego i testowego za pomocą Partitioning w odpowiednich proporcjach.
4. Połączenie czterech zbiorów uczących w jeden i połączenie czterech zbiorów testowych w jeden za pomocą Concatenate.
5. Stworzenie trzech modeli na zbiorze uczącym za pomocą węzłów Decision Tree Learner, Gradient Boosted Trees Learner i Random Forest Learner.
6. Testowanie modeli za pomocą węzłów Decision Tree Predictor, Gradient Boosted Trees Predictor i Random Forest Predictor.
7. Ewaluacja modeli za pomocą węzła Scorer.

Wyniki dla poszczególnych modeli zostały przedstawione poniżej:

model\podział	70:30	80:20	90:10																																																																											
Drzewo decyzyjne	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>34</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>54</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>30</td><td>0</td></tr><tr><th>3</th><td>1</td><td>0</td><td>0</td><td>34</td></tr></table> <p>Correct classified: Wrong classified: 1 Accuracy: 99,346% Error: 0,654% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	34	0	0	0	4	0	54	0	0	1	0	0	30	0	3	1	0	0	34	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>23</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>36</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>20</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>23</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	23	0	0	0	4	0	36	0	0	1	0	0	20	0	3	0	0	0	23	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>12</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>18</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>10</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>12</td></tr></table> <p>Correct classified: 52 Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	12	0	0	0	4	0	18	0	0	1	0	0	10	0	3	0	0	0	12
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	34	0	0	0																																																																										
4	0	54	0	0																																																																										
1	0	0	30	0																																																																										
3	1	0	0	34																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	23	0	0	0																																																																										
4	0	36	0	0																																																																										
1	0	0	20	0																																																																										
3	0	0	0	23																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	12	0	0	0																																																																										
4	0	18	0	0																																																																										
1	0	0	10	0																																																																										
3	0	0	0	12																																																																										
Gradient Boosted	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>34</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>54</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>30</td><td>0</td></tr><tr><th>3</th><td>1</td><td>0</td><td>0</td><td>34</td></tr></table> <p>Correct classified: Wrong classified: 1 Accuracy: 99,346% Error: 0,654% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	34	0	0	0	4	0	54	0	0	1	0	0	30	0	3	1	0	0	34	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>23</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>36</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>20</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>23</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	23	0	0	0	4	0	36	0	0	1	0	0	20	0	3	0	0	0	23	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>12</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>18</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>10</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>12</td></tr></table> <p>Correct classified: 52 Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	12	0	0	0	4	0	18	0	0	1	0	0	10	0	3	0	0	0	12
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	34	0	0	0																																																																										
4	0	54	0	0																																																																										
1	0	0	30	0																																																																										
3	1	0	0	34																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	23	0	0	0																																																																										
4	0	36	0	0																																																																										
1	0	0	20	0																																																																										
3	0	0	0	23																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	12	0	0	0																																																																										
4	0	18	0	0																																																																										
1	0	0	10	0																																																																										
3	0	0	0	12																																																																										
Las losowy	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>34</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>54</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>30</td><td>0</td></tr><tr><th>3</th><td>1</td><td>0</td><td>0</td><td>34</td></tr></table> <p>Correct classified: Wrong classified: 1 Accuracy: 99,346% Error: 0,654% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	34	0	0	0	4	0	54	0	0	1	0	0	30	0	3	1	0	0	34	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>23</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>36</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>20</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>23</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	23	0	0	0	4	0	36	0	0	1	0	0	20	0	3	0	0	0	23	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>12</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>18</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>10</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>12</td></tr></table> <p>Correct classified: 52 Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	12	0	0	0	4	0	18	0	0	1	0	0	10	0	3	0	0	0	12
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	34	0	0	0																																																																										
4	0	54	0	0																																																																										
1	0	0	30	0																																																																										
3	1	0	0	34																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	23	0	0	0																																																																										
4	0	36	0	0																																																																										
1	0	0	20	0																																																																										
3	0	0	0	23																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	12	0	0	0																																																																										
4	0	18	0	0																																																																										
1	0	0	10	0																																																																										
3	0	0	0	12																																																																										

Na podstawie przedstawionych wyników można zauważyć, że różne proporcje podziału danych na zbiory uczące i testowe (70:30, 80:20, 90:10) mają niewielki wpływ na skuteczność zastosowanych modeli. Wszystkie trzy algorytmy – Drzewo Decyzyjne, Gradient Boosted oraz Las Losowy – wykazują bardzo wysoką dokładność. Modele uzyskują niemal idealne wyniki, co potwierdzają zarówno dokładność, jak i wartość wskaźnika Cohen's kappa.

W przypadku drzewa decyzyjnego dokładność przy podziale 70:30 wynosi 99,35%, co oznacza, że tylko jedna próba została błędnie sklasyfikowana. Wartość wskaźnika Cohen's kappa na poziomie bliskim 1 wskazuje na bardzo wysoką zgodność predykcji modelu. Przy proporcjach 80:20 i 90:10 model osiąga już 100% dokładności – wszystkie przypadki zostały poprawnie sklasyfikowane, a wskaźnik błędów wynosi 0%.

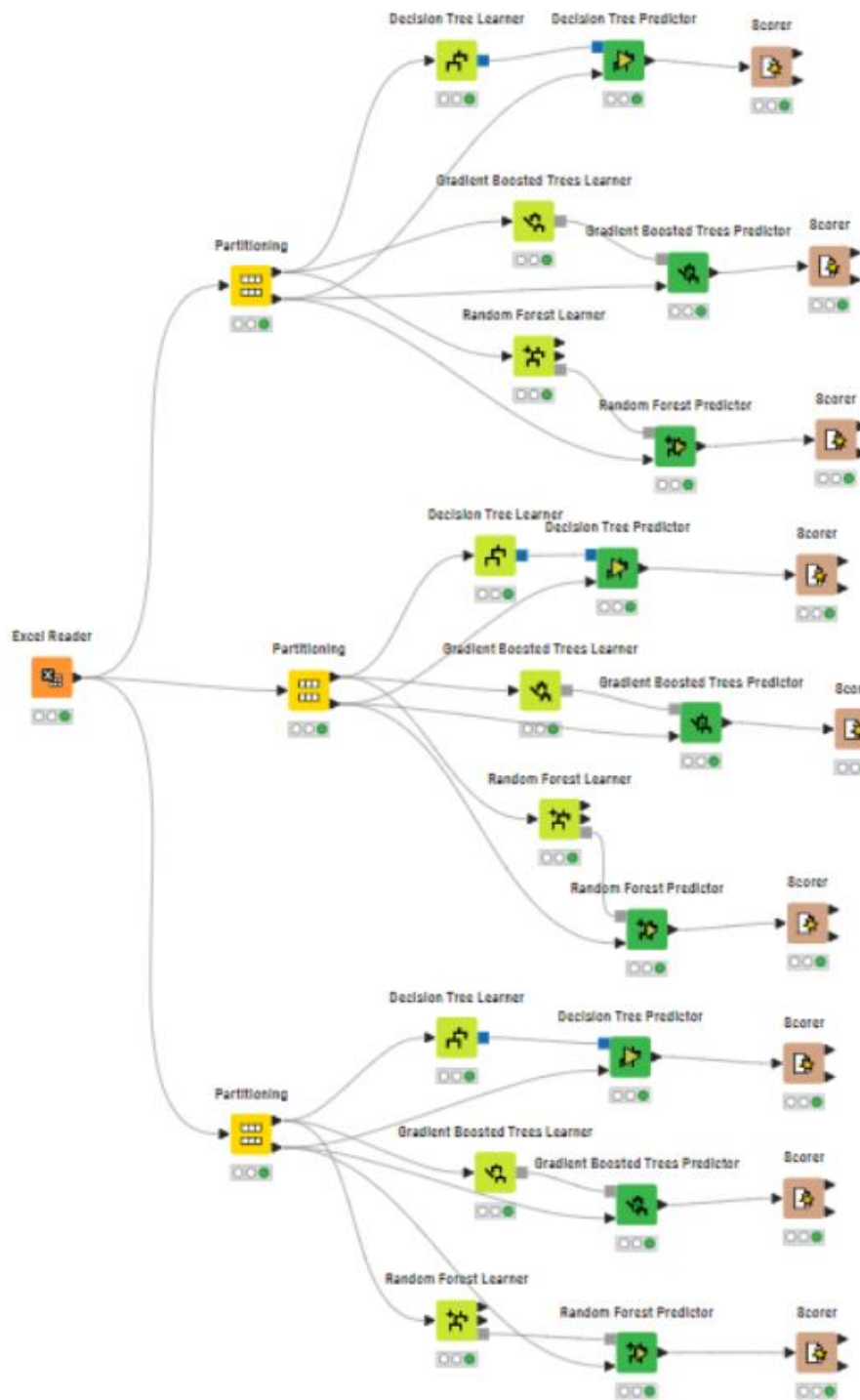
Podobne wyniki uzyskano dla modelu Gradient Boosted. Przy podziale 70:30, dokładność wynosi 99,35%, co również oznacza tylko jeden błąd klasyfikacji. Przy proporcji 80:20 jak i 90:10 skuteczność modelu wzrasta do 100%, a więc błędów klasyfikacji nie występują. To wskazuje, że algorytm radzi sobie równie dobrze, jak drzewo decyzyjne w tych warunkach.

Las losowy również przy podziale 70:30 wykazuje dokładność 99,35%, a liczba błędnie sklasyfikowanych prób to 1. Przy podziale jak 80:20 i 90:10 dokładność wzrasta do 100%, a model klasyfikuje wszystkie przypadki poprawnie. Wartości Cohen's kappa dla wszystkich trzech modeli są niemal idealne, co potwierdza ich skuteczność.

Można również zauważyć, że zwiększenie proporcji danych uczących (np. 80:20) prowadzi do lepszych wyników – model posiada więcej danych do nauki, co pozwala na dokładniejsze klasyfikowanie. Mimo to różnice pomiędzy proporcjami 80:20 i 90:10 są minimalne. Wszystkie trzy modele wydają się stabilne i dobrze dopasowane do problemu, skoro różnice w wynikach są minimalne. Zastosowanie bardziej zaawansowanych algorytmów, takich jak Gradient Boosted czy Las Losowy, nie przyniosło widocznych korzyści w porównaniu do prostszego drzewa decyzyjnego.

Podział bezpośredni:

Podejście z bezpośrednim podziałem danych zostało natomiast przedstawione na poniższym rysunku.



Budowa modeli różni się od poprzedniego podejścia jedynie brakiem wyodrębniania grup.

W tym przypadku dostajemy następujące wyniki:

model\po dział	70:30	80:20	90:10																																																																											
Drzewo decyzyjne	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>30</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>66</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>16</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>38</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	30	0	0	0	4	0	66	0	0	1	0	0	16	0	3	0	0	0	38	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>21</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>27</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>22</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>30</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	21	0	0	0	4	0	27	0	0	1	0	0	22	0	3	0	0	0	30	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>13</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>17</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>9</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>11</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	13	0	0	0	4	0	17	0	0	1	0	0	9	0	3	0	0	0	11
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	30	0	0	0																																																																										
4	0	66	0	0																																																																										
1	0	0	16	0																																																																										
3	0	0	0	38																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	21	0	0	0																																																																										
4	0	27	0	0																																																																										
1	0	0	22	0																																																																										
3	0	0	0	30																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	13	0	0	0																																																																										
4	0	17	0	0																																																																										
1	0	0	9	0																																																																										
3	0	0	0	11																																																																										
Gradient Boosted	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>30</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>66</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>16</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>38</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	30	0	0	0	4	0	66	0	0	1	0	0	16	0	3	0	0	0	38	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>21</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>27</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>22</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>30</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	21	0	0	0	4	0	27	0	0	1	0	0	22	0	3	0	0	0	30	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>13</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>17</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>9</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>11</td></tr></table> <p>Correct classified: 50 Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	13	0	0	0	4	0	17	0	0	1	0	0	9	0	3	0	0	0	11
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	30	0	0	0																																																																										
4	0	66	0	0																																																																										
1	0	0	16	0																																																																										
3	0	0	0	38																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	21	0	0	0																																																																										
4	0	27	0	0																																																																										
1	0	0	22	0																																																																										
3	0	0	0	30																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	13	0	0	0																																																																										
4	0	17	0	0																																																																										
1	0	0	9	0																																																																										
3	0	0	0	11																																																																										
Las losowy	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>30</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>66</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>16</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>38</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	30	0	0	0	4	0	66	0	0	1	0	0	16	0	3	0	0	0	38	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>21</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>27</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>22</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>30</td></tr></table> <p>Correct classified: Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	21	0	0	0	4	0	27	0	0	1	0	0	22	0	3	0	0	0	30	<table><tr><th>grupa \ Prediction (grupa)</th><th>2</th><th>4</th><th>1</th><th>3</th></tr><tr><th>2</th><td>13</td><td>0</td><td>0</td><td>0</td></tr><tr><th>4</th><td>0</td><td>17</td><td>0</td><td>0</td></tr><tr><th>1</th><td>0</td><td>0</td><td>9</td><td>0</td></tr><tr><th>3</th><td>0</td><td>0</td><td>0</td><td>11</td></tr></table> <p>Correct classified: 50 Wrong classified: 0 Accuracy: 100% Error: 0% Cohen's kappa (κ):</p>	grupa \ Prediction (grupa)	2	4	1	3	2	13	0	0	0	4	0	17	0	0	1	0	0	9	0	3	0	0	0	11
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	30	0	0	0																																																																										
4	0	66	0	0																																																																										
1	0	0	16	0																																																																										
3	0	0	0	38																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	21	0	0	0																																																																										
4	0	27	0	0																																																																										
1	0	0	22	0																																																																										
3	0	0	0	30																																																																										
grupa \ Prediction (grupa)	2	4	1	3																																																																										
2	13	0	0	0																																																																										
4	0	17	0	0																																																																										
1	0	0	9	0																																																																										
3	0	0	0	11																																																																										

Wszystkie trzy modele – Drzewo Decyzyjne, Gradient Boosted i Las Losowy – osiągają idealną skuteczność we wszystkich proporcjach podziału danych na zbiory uczące i testowe (70:30, 80:20, 90:10). Dokładność wynosi w każdym przypadku 100%, brak jest błędów klasyfikacji, a wskaźnik Cohen's kappa ma wartość 1, co świadczy o doskonałej zgodności przewidywań modeli z rzeczywistymi klasami.

Przy podziale 70:30 każdy model poprawnie sklasyfikował wszystkie 30 próbek w zbiorze testowym. Podział 80:20 daje równie idealny wynik, gdzie 20 próbek zostało poprawnie sklasyfikowanych. Dla największego podziału zbioru uczącego (90:10) modele również klasyfikują bezbłędnie, poprawnie przypisując wszystkie 10 próbek w zbiorze testowym do odpowiednich klas.

Taki wynik pokazuje, że wszystkie trzy algorytmy radzą sobie równie dobrze z tym zadaniem klasyfikacyjnym. Wszystkie zastosowane algorytmy skutecznie wychwytyją wzorce w danych, niezależnie od ilości próbek użytych do treningu.

Porównanie wyników przy podziale wieloetapowym (stratyfikacją) i bezpośrednim:

Porównując wyniki ze stratyfikacją i bez niej, widać, że przy stratyfikacji modele popełniały pojedynczy błąd dla proporcji 70:30, osiągając dokładność 99,35%, natomiast dla 80:20 90:10 wyniki były idealne. Bez stratyfikacji wszystkie modele uzyskały 100% dokładności we wszystkich proporcjach, niezależnie od wielkości zbioru uczącego. Sugeruje to, że w tym przypadku dane były na tyle dobrze zbalansowane i łatwe do klasyfikacji, że brak stratyfikacji nie wpłynął negatywnie na skuteczność modeli, a wręcz pozwolił osiągnąć perfekcyjne wyniki.

Porównując wyniki uzyskane przy podziale danych ze stratyfikacją i bez niej, widać wyraźne różnice w wiarygodności oceny jakości modeli. W podejściu ze stratyfikacją proporcje klas były zachowane w zbiorach uczących i testowych, co zapewniało bardziej reprezentatywne warunki do ewaluacji. Modele osiągały bardzo wysoką dokładność, z drobnymi błędami klasyfikacyjnymi, co wskazuje na ich zdolność do generalizacji w bardziej zróżnicowanym zbiorze testowym. Nawet te niewielkie błędy pokazują, że modele rzeczywiście uczą się rozpoznawania wzorców w danych, a nie tylko zapamiętują specyfikę zbioru uczącego.

Z kolei brak stratyfikacji w podziale danych prowadził do sytuacji, w której wszystkie modele osiągnęły idealną skuteczność – 100% dokładności, niezależnie od wielkości podziału na zbiór uczący i testowy. Choć na pierwszy rzut oka takie wyniki mogą wydawać się doskonałe, w rzeczywistości budzą wątpliwości. Brak zachowania proporcji klas w zbiorach mógł prowadzić do nierównomiernego rozkładu danych, gdzie zbiór testowy był albo zbyt podobny do danych uczących, albo nie zawierał wystarczającej różnorodności. W efekcie modele mogły uzyskiwać perfekcyjne wyniki nie dlatego, że dobrze generalizowały, ale dlatego, że ich zadanie było uproszczone przez przypadkowy charakter danych testowych.

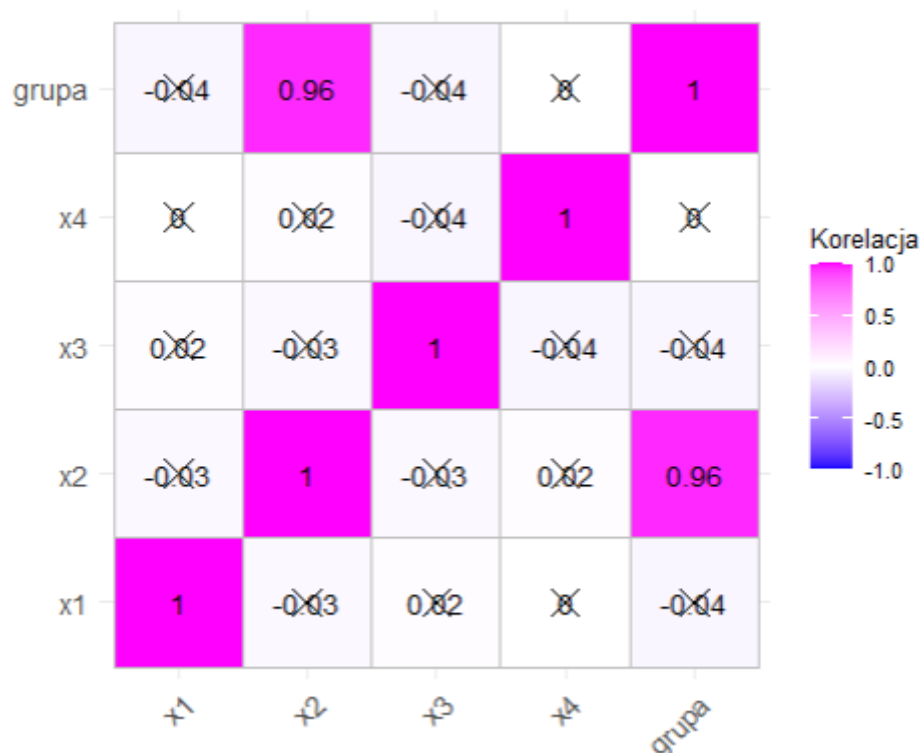
Przy podziale 90:10 problem ten jest szczególnie widoczny. Mały zbiór testowy zawiera bardzo niewiele próbek, co zmniejsza różnorodność problemów, z jakimi modele muszą się zmierzyć. W takich warunkach idealna skuteczność może wynikać z dopasowania do specyfiki danych, a nie z rzeczywistej zdolności modeli do radzenia sobie z nowymi, niezależnymi danymi.

Podsumowując, podejście ze stratyfikacją okazuje się bardziej wiarygodne i daje lepszy obraz rzeczywistej skuteczności modeli. Stratyfikacja wymusza równowagę między klasami w zbiorach uczących i testowych, co stawia przed modelami większe wyzwanie i pozwala ocenić ich zdolność do generalizacji. Wyniki uzyskane bez stratyfikacji, choć imponujące, należy traktować z dużą ostrożnością, ponieważ mogą być wynikiem przypadkowego dopasowania, a nie rzeczywistej efektywności.

Badanie istotności zmiennych

W ramach dalszej analizy zostanie przeprowadzona identyfikacja zmiennych odgrywających kluczową rolę w budowie modeli klasyfikacyjnych.

W celu wyznaczenia istotności zmiennych przy podziale na grupy posłużymy się analizą macierzy korelacji. Narzędzie to pozwala ocenić siłę i kierunek zależności pomiędzy zmiennymi, co umożliwi identyfikację czynników kluczowych dla procesu grupowania. Wartości współczynnika korelacji Spearman'a zostały zilustrowane zarówno liczbowo, jak i graficznie – różowy kolor oznacza silną dodatnią korelację, biały brak zależności, a niebieski silną ujemną korelację.



Analiza macierzy korelacji wskazuje, że zmienna **x2** jest bardzo silnie skorelowana ze zmienną „grupa” ($R = 0,96$), co czyni ją najistotniejszą zmienną w kontekście podziału na grupy. Pozostałe zmienne, takie jak **x1**, **x3** czy **x4**, nie wykazują istotnych korelacji ze zmienną „grupa”, co potwierdza ich znikome znaczenie w procesie grupowania. Ich współczynniki korelacji są na tyle niskie, że można je uznać za nieistotne dla analizy, co zostało dodatkowo oznaczone w macierzy przekreśleniami.

Wyniki jednoznacznie wskazują, że zmienna **x2** odgrywa kluczową rolę w dalszej analizie.

Macierz korelacji została opracowana w języku R przy wykorzystaniu bibliotek `dplyr` oraz `ggcorrplot`. Poniżej znajduje się kod użyty do jej wygenerowania:


```

library(dplyr)
library(ggcorrplot)

# Wybór tylko kolumn numerycznych
Correlation_Variables_numeric <- dane %>%
  select_if(is.numeric)

# Obliczenie macierzy korelacji metoda Spearmana
Correlation_matrix_numeric <- round(cor(Correlation_Variables_numeric, method = "spearman"), 3)

# Obliczenie macierzy p-wartości
p.mat_coefficient <- cor_pmat(Correlation_Variables_numeric)

# Wizualizacja z wyświetlaniem wszystkich korelacji
ggcorrplot(
  Correlation_matrix_numeric,
  lab = TRUE, # Wyświetlanie wartości korelacji
  p.mat = p.mat_coefficient, # Macierz p-wartości
  sig.level = 0.05, # Poziom istotności
  colors = c("blue", "white", "magenta"), # Gradient kolorów
  legend.title = "Korelacja"
)

```

Ze względu na różnorodność dostępnych metod, przeprowadzono analizę istotności zmiennych na podstawie modeli klasyfikacyjnych, takich jak drzewa decyzyjne i lasy losowe. W trakcie badań zastosowano dwa podejścia do podziału danych: ze stratyfikacją oraz bez niej, aby lepiej zrozumieć wpływ zmiennych w różnych warunkach. Niestety, dla modelu gradient boosting w KNIME nie było możliwości odczytania wyników istotności zmiennych w sposób analogiczny do pozostałych metod, dlatego analiza została ograniczona do drzew decyzyjnych i lasów losowych.

Podejście ze stratyfikacją:

We wszystkich wariantach podziału danych na zbiór uczący i testowy (70:30, 80:20 oraz 90:10), otrzymane drzewa decyzyjne wykonały cztery podziały w celu klasyfikacji danych (zdjęcia: drzewo1.png, drzewo2.png, drzewo3.png). Wszystkie te podziały opierały się na wartości zmiennej **x2**, co jednoznacznie wskazuje na jej dominującą istotność w analizowanym zbiorze.

W przypadku lasów losowych, klasyfikacja opiera się na agregacji wyników uzyskanych z wielu drzew decyzyjnych, które są tworzone na podstawie losowych podzbiorów danych i zmiennych.

Przy podziale zbioru uczącego i testowego w proporcjach 70:30 (zdjęcie: las1.png) zmienne **x2** i **x4** odgrywają kluczową rolę w budowie lasu losowego, ponieważ to one najczęściej pojawiają się jako punkty podziałów w drzewach. Z danych uzyskanych z lasu losowego wynika, że podziały oparte na zmiennej **x2** miały decydujący wpływ na przypisanie próbek do odpowiednich klas (wystąpiła ona aż 3 razy spośród 4 podziałów).

W przypadku podziału 80:20 (zdjęcie: las2.png) widać, że zmienne takie jak **x1** i **x2** mają największy wpływ na klasyfikację, pojawiając się na kluczowych poziomach drzewa. Ich obecność w węzłach blisko korzenia wskazuje, że decydują o głównych podziałach w

danych. Z kolei zmienne **x3** i **x4** pojawiają się głównie w niższych węzłach, co sugeruje ich mniejszą istotność w ogólnej strukturze modelu, chociaż nadal wpływają na lokalne podziały.

Przy podziale 90:10 (zdjęcie: las3.png) można zauważyć, że zmienna **x1** odgrywa kluczową rolę w klasyfikacji, pojawiając się w korzeniu drzewa. Jej obecność w tym węźle wskazuje, że decyduje o głównym podziale w danych, co czyni ją najbardziej istotną zmienną w modelu. Na kolejnych poziomach, w węzłach blisko korzenia, istotny wpływ mają również zmienne **x2** i **x3**, które odpowiadają za kolejne ważne podziały. Ich obecność na wyższych poziomach drzewa sugeruje ich wysoką istotność w wyjaśnianiu zmienności danych, lecz nie aż tak wysoką jak **x1**. Z kolei zmienna **x4** pojawiająca się w niższych węzłach drzewa ma mniejszy wpływ na ogólną strukturę modelu, ale nadal odgrywają istotną rolę w lokalnych podziałach.

Podejście z bezpośrednim podziałem danych:

We wszystkich przeprowadzonych podziałach danych na zbiory uczący i testowy (70:30, 80:20 oraz 90:10), uzyskane drzewa decyzyjne klasyfikowały dane, dokonując czterech podziałów w każdym przypadku (wizualizacje: drzewo4.png, drzewo5.png, drzewo6.png). Wszystkie te podziały bazowały na wartości zmiennej **x2**, co wyraźnie wskazuje na jej kluczowe znaczenie w analizowanym zestawie danych.

Przy podziale 70:30 (zdjęcie: las4.png) można zauważyć, że zmienna **x2** odgrywa kluczową rolę w klasyfikacji, pojawiając się w korzeniu drzewa. Jej obecność w tym węźle wskazuje, że decyduje o głównym podziale w danych, co czyni ją najbardziej istotną zmienną w modelu. Na kolejnych poziomach, w węzłach blisko korzenia, istotny wpływ mają również zmienne **x1** i **x3**, które odpowiadają za kolejne ważne podziały. Ich obecność na wyższych poziomach drzewa sugeruje ich wysoką istotność w wyjaśnianiu zmienności danych, lecz nie aż tak wysoką jak **x2**.

W przypadku podziału 80:20 (zdjęcie: las5.png) zmienna **x1** odgrywa kluczową rolę w klasyfikacji, pojawiając się na korzeniu drzewa i decydując o głównym podziale danych. Na kolejnych poziomach istotnymi zmiennymi są **x3** i **x4**, które wpływają na dalsze kluczowe podziały, zapewniając precyzję modelu. Zmienna **x2** pojawia się głównie w niższych węzłach, co wskazuje na jej mniejszą istotność w ogólnej strukturze drzewa, choć ma lokalne znaczenie w doprecyzowaniu klasyfikacji. Taka hierarchia zmiennych sugeruje, że **x1** ma największą moc separacyjną.

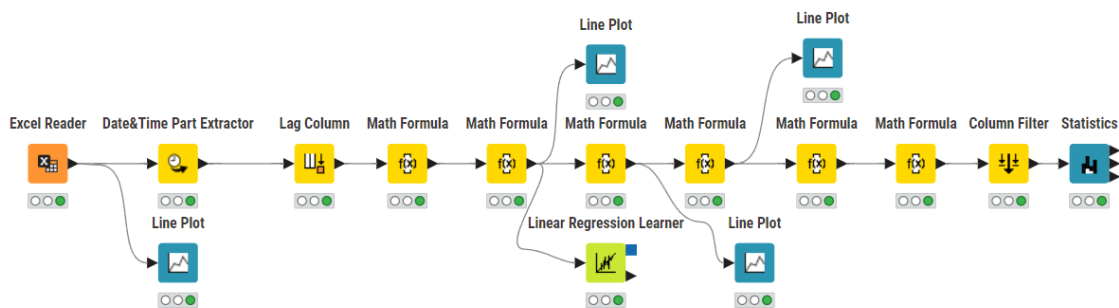
Na podstawie struktury lasu losowego (zdjęcie: las6.png) przy podziale zbioru w stosunku 90:10 najbardziej istotną zmienną jest **x2**, ponieważ jest używana w korzeniu drzewa i od niej zaczynają się kluczowe podziały, wpływając na największą liczbę obserwacji. Kolejną istotną zmienną jest **x4**, która pojawia się wielokrotnie na niższych poziomach drzewa, co świadczy o jej znaczącej roli w dalszym różnicowaniu danych.

Zmienna **x1** również odgrywa ważną rolę, choć w mniejszym stopniu, pojawiając się w specyficznych podziałach na niższych poziomach.

Podsumowując, analiza wskazuje, że zmienna **x2** odgrywa dominującą rolę w klasyfikacji danych w większości modeli, zwłaszcza w drzewach decyzyjnych, gdzie konsekwentnie pojawia się w korzeniu drzewa. Hierarchia istotności zmiennych w lasach losowych różni się w zależności od proporcji podziału danych.

Zadanie 2.

Celem drugiego zadania było stworzenie modelu opartego na prostej regresji liniowej w celu analizy trendów dotyczących obrotów oraz ocena efektywności modelu na podstawie analizy błędów względnych. W ramach zadania uwzględniony został wpływ skoków sezonowych zależnych od dnia tygodnia.

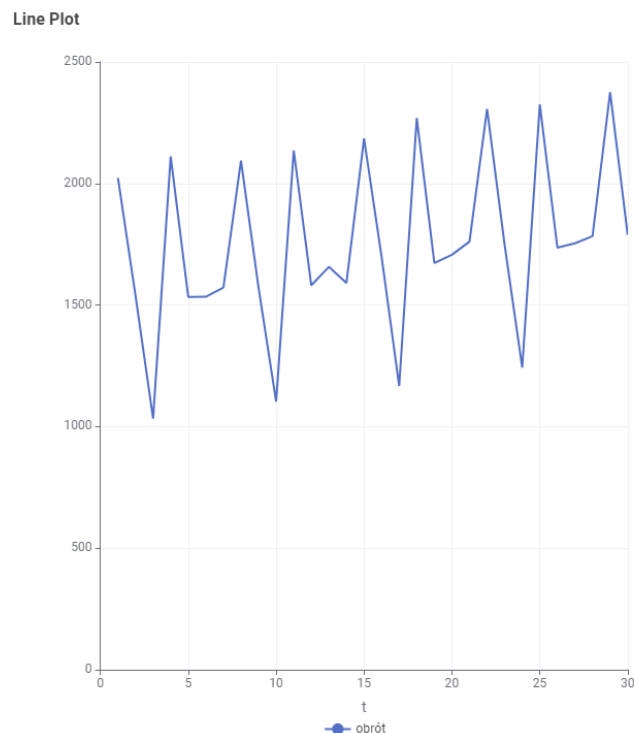


Workflow w KNIME

Opis wykonanych działań:

1. Wczytywanie danych

Dane wczytane z pliku prognozowanie1.xlsx za pomocą węzła Excel Reader.



Wykres obrotu do czasu (pierwsze 30 dni)

2. Ekstrakcja cech czasowych

Za pomocą węzła Date&Time Part Extractor wyciągnięto numer dnia tygodnia, do użycia przy analizie wpływu dnia tygodnia na obrót.

3. Opóźnienie czasowe

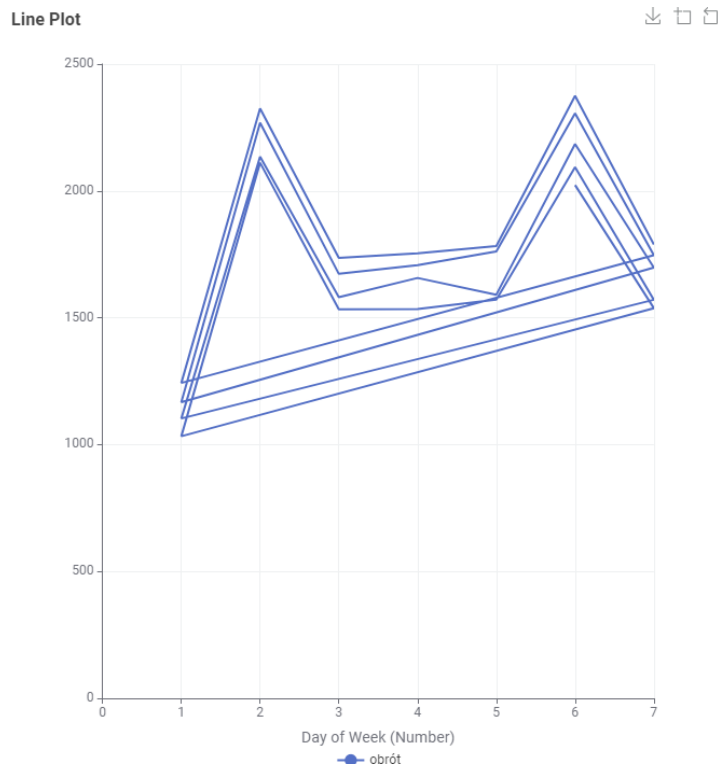
Za pomocą węzła Lag Column utworzono nową kolumnę “obróć(-1)”, zawierającą wartość obrotu z dnia poprzedniego. Wprowadzenie takich zależności czasowych pozwala na uwzględnienie wpływu zmian w przeszłości na teraźniejszość w analizie.

4. Obliczenie różnic w obrotach

$\$obróć - \$obróć(-1)$

Za pomocą węzła Math Formula obliczono różnicę między wartością obrotu z dnia bieżącego a wartością z dnia poprzedniego.

5. Modelowanie skoków sezonowych



Z rysunku można zauważyć wyraźne zmiany w poniedziałki, wtorki i soboty. Dokładne wahania w tych dniach można odczytać z kolumny \$roznica\$.

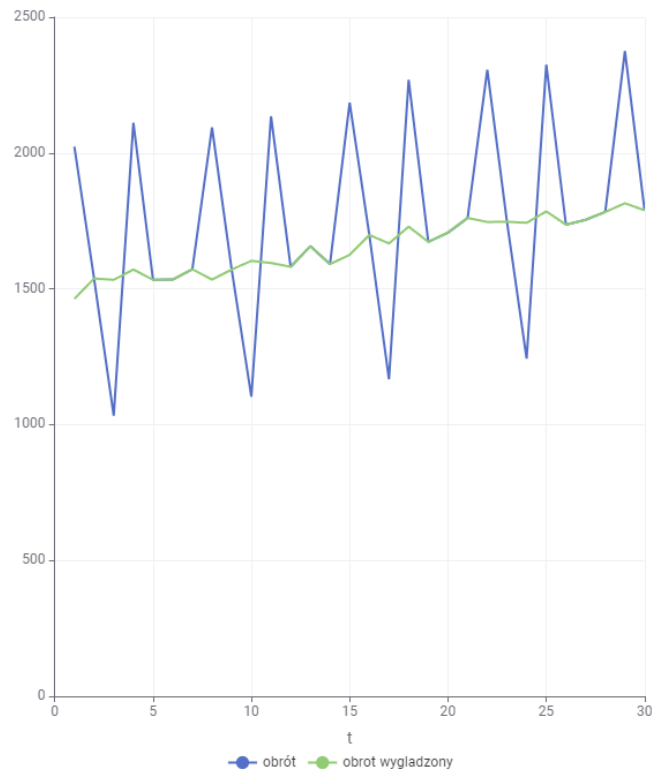
Przy użyciu instrukcji warunkowych dokonano korekty wartości obrotu w konkretnych dniach tygodnia, aby uwzględnić specyficzne odchylenia obrotu od dnia tygodnia. Wartości obrotu po korekcie zostały zapisane w nowej kolumnie “obróć skorygowany”.

```

1 if($Day of Week (Number)$ == 1, $obrót$+500,
2 if($Day of Week (Number)$ == 2, $obrót$-540,
3 if($Day of Week (Number)$ == 6, $obrót$-560,$obrót$)))

```

Line Plot



Porównanie obrotu do obrotu po wygładzeniu skoków sezonowych (pierwsze 30 dni)

6. Model regresji liniowej

Variable String	Coef. Number (double)	Std. Err. Number (double)	t-value Number (double)	P> t Number (double)
t	10.006	0.005	2,098.487	0
Intercept	1,497.675	2.012	744.498	0

Współczynniki i statystyki modelu regresji liniowej

Za pomocą węzła Linear Regression Learner zbudowano model regresji liniowej na podstawie obrotów skorygowanych i wyznaczono równanie prostej opisującej trend czasowy.

$$10 * t + 1500$$

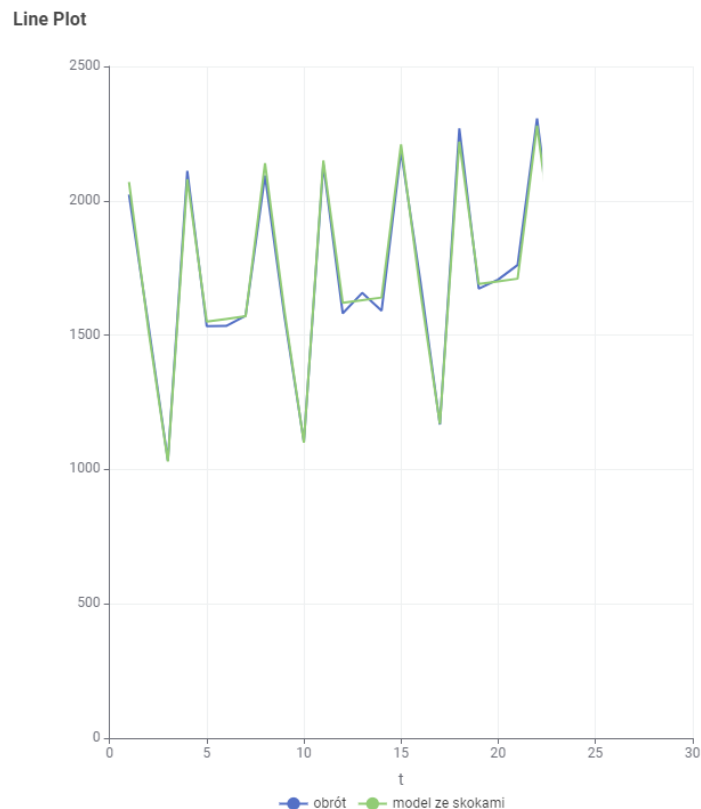
7. Rekonstrukcja skoków sezonowych

```

1 if($Day of Week (Number)$ == 1, $model$-500,
2 if($Day of Week (Number)$ == 2, $model$+540,
3 if($Day of Week (Number)$ == 6, $model$+560,$model$)

```

Ponownie dodano skoki sezonowe, aby utworzyć pełny model wartości obrotów



Porównanie obrotu do modelu z dodanymi skokami sezonowymi (pierwsze 30 dni)

8. Obliczenie wartości błędów

- Błąd bezwzględny

```
abs($obróć-$model ze skokami$)
```

- Błąd względny

```
$błąd bezwzględny/$obróć$
```

9. Analiza statystyczna błędów

Column String	Min Number (dou...)	Max Number (dou...)	Mean Number (dou...)	Std. devi... Number (dou...)	Variance Number (dou...)	Skewness Number (dou...)	Kurtosis Number (dou...)
błąd względny	0	0.031	0.005	0.005	0	1.665	3.634

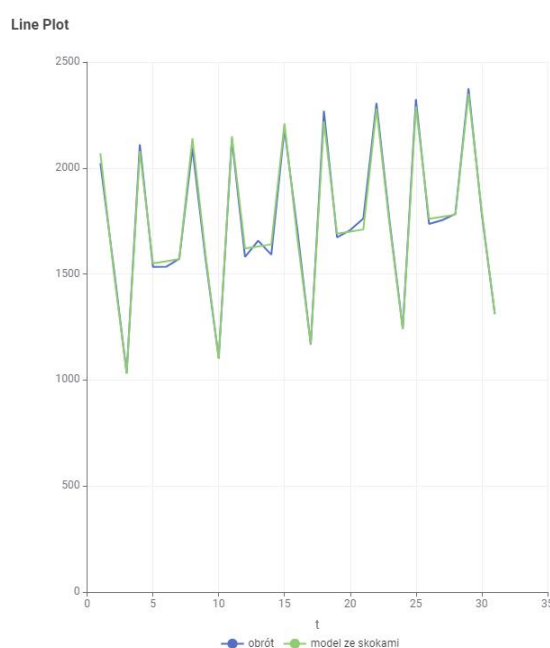
- Minimalny błąd względny wynosi 0, co oznacza, że dla niektórych obserwacji model osiągnął idealne dopasowanie.
- Maksymalny błąd względny wynosi 0,0308, co wskazuje, że największe odchylenie prognozy od rzeczywistej wartości jest niewielkie.
- Średnia wartość błędu względnego wynosi 0,005, co świadczy o ogólnej dokładności modelu – przeciętne odchylenie od rzeczywistości jest małe.
- Odchylenie standardowe wynosi 0,0048, co oznacza, że błąd względny ma niewielkie wahania wokół średniej, a model jest stabilny.
- Wartość skośności wynosi 1,6654, co sugeruje, że rozkład błędów jest prawostronnie skośny. Większość błędów względnych jest bliska średniej,

ale istnieje niewielka liczba większych odchyłek w kierunku wyższych wartości błędów.

- Wartość kurtozy wynosi 3,6344, co oznacza, że rozkład ma bardziej ostro zakończone wierzchołki w porównaniu do rozkładu normalnego. Może to wskazywać na obecność większej liczby obserwacji bliskich średniej oraz kilku odchyłek skrajnych.

10. Wizualizacja wyników

- Korzystając z węzła Line Plot zwizualizowano rzeczywiste i modelowane wartości obrotów w celu oceny jakości modelu na podstawie przedstawienia graficznego wyników.



Podsumowanie:

Dopasowanie trendu: Model regresji liniowej z równaniem $10 \cdot t + 1500$, po uwzględnieniu skoków sezonowych, dobrze opisuje ogólny trend w danych.

Wnioski:

1. Efektywność modelu liniowego

- Model regresji liniowej, oparty na trendach czasowych i uwzględniający skoki sezonowe, skutecznie odzwierciedla ogólny wzorzec zmian w danych. Dzięki zastosowaniu funkcji liniowej oraz korekt sezonowych osiągnięto stabilne wyniki z niskim średnim błędem względnym wynoszącym 0,5%.

2. Znaczenie korekty sezonowej

- Uwzględnienie specyficznych zmian zależnych od dnia tygodnia znacząco poprawiło jakość predykcji. Umożliwiło to lepsze dopasowanie modelu do rzeczywistych danych, co wskazuje na istotność analizy sezonowości w predykcji.

3. Błędy modelu

- Średni błąd względny 0,005 jest stosunków mały, co świadczy o dobrej dokładności modelu.
- Rozkład błędów jest stabilny, a ich odchylenie standardowe wynosiło 0,5%.
- Ekstrema błędów (0 i 0,031) wskazują na dobrą precyzję modelu, choć mogą występować niewielkie odchylenia w specyficznych punktach danych.

4. Skalowalność rozwiązania

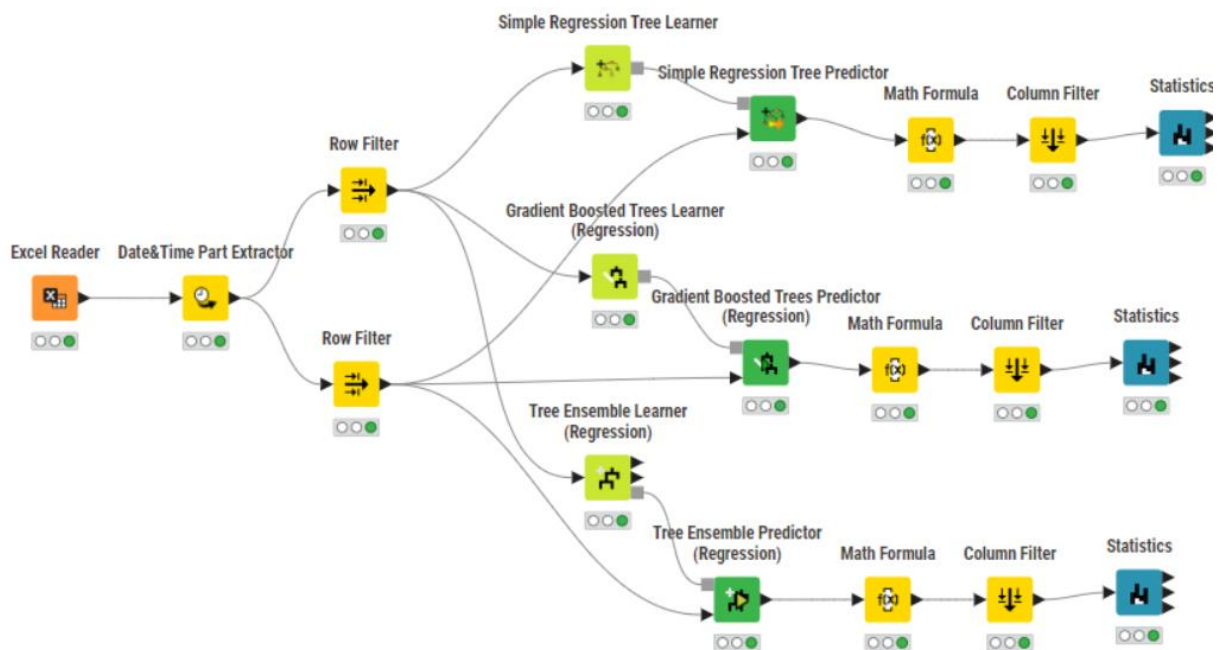
- Model jest stosunkowo prosty, co pozwala na jego szybkie wdrożenie w praktyce, ale w przypadku bardziej złożonych danych warto rozważyć rozszerzenie analizy o bardziej zaawansowane techniki modelowania

Zadanie 3.

W zadaniu tym wybrano ostatni rok jako zbiór testowy, a następnie zbudowano trzy różne modele uczenia maszynowego:

- **Proste drzewo regresyjne** - w drzewie tym przewidywana wartość dla węzła liścia jest średnią wartością docelową rekordów w liściu. W związku z tym przewidywania są najlepsze (w odniesieniu do danych szkoleniowych), jeśli wariancja wartości docelowych w liściu jest minimalna. Osiąga się to poprzez podziały, które minimalizują sumę kwadratów błędów w ich odpowiednich elementach potomnych.
- **Gradient Boosted Trees** - opisane w zadaniu 1,
- **Zespół drzew (Tree Ensemble)** - wykorzystuje wiele drzew decyzyjnych do przewidywania wyników. Każde drzewo jest budowane na podstawie różnych podzbiorów danych treningowych i/lub atrybutów. W przypadku regresji, przewidywana wartość dla węzła liścia to średnia wartość docelowa rekordów w liściu. Model taki łączy prognozy z każdego drzewa, najczęściej przy użyciu prostej średniej, aby uzyskać bardziej ogólną i dokładniejszą predykcję. Zespół drzew minimalizuje błąd prognozy poprzez ograniczenie wariancji wartości docelowych w liściach.

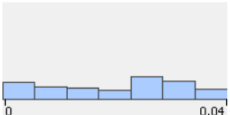
Modele te przedstawiono na poniższym rysunku.



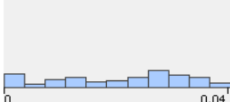
Poszczególne etapy budowy modeli:

1. Wczytanie danych.
2. Wyodrębnienie z danych z kolumny data informacji o roku za pomocą węzła Date&Time Part Extractor.
3. Wyfiltrowanie danych z ostatniego roku jako zbiór testowy (pozostałe dane jako zbiór uczący) za pomocą węzłów Row Filter.
4. Budowa różnych modeli za pomocą węzłów Simple Regression Tree Learner, Gradient Boosted Trees Learner (Regression) i Tree Ensemble Learner (Regression).
5. Testowanie modeli za pomocą węzłów Simple Regression Tree Predictor, Gradient Boosted Trees Predictor (Regression) i Tree Ensemble Predictor (Regression).
6. Obliczenie błędu względnego za pomocą węzłów Math Formula jako wartości bezwzględnej z różnicy między prawdziwym obrotem a wynikiem predykcji, podzielonej przez prawdziwy obrót.
7. Wyodrębnienie z danych kolumny o błędzie względnym za pomocą węzła Column Filter (w celu prawidłowego działania następnego węzła).
8. Ewaluacja modeli na podstawie błędu względnego za pomocą węzła Statistic.

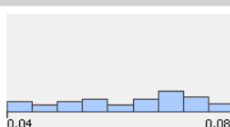
Wyniki ewaluacji dla pojedynczego drzewa:

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
błąd względny	7,54E-6	0,0206	0,0238	0,0414	0,012	-0,2142	-1,2566	0	0	0	

Wyniki ewaluacji dla Gradient Boosted:

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
błąd względny	0,0001	0,0223	0,0258	0,0434	0,0125	-0,3013	-1,1631	0	0	0	

Wyniki ewaluacji dla zespołu drzew:

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
błąd względny	0,0382	0,0623	0,0658	0,0827	0,0123	-0,3684	-1,0544	0	0	0	

Dla pojedynczego drzewa średni błąd względny wynosi 0,0206, co czyni go najmniejszym spośród wszystkich modeli. Drzewo to cechuje się też stosunkowo małą medianą (0,0238) i maksymalnym błędem (0,0414), co wskazuje na jego dobrą dokładność. Rozkład błędów ma lekką asymetrię w stronę wartości ujemnych (skośność -0,2142) oraz płaską charakterystykę (kurtosis -1,2566).

W przypadku Gradient Boosted Trees średni błąd względny jest nieco wyższy i wynosi 0,0223, a mediana i maksymalny błąd to odpowiednio 0,0258 i 0,0434. Odchylenie standardowe (0,0125) wskazuje na stabilność błędów, a rozkład błędów jest bardziej asymetryczny (-0,3013) w porównaniu do pojedynczego drzewa. Gradient Boosted Trees oferują jednak bardziej zaawansowane podejście, pozwalające lepiej modelować złożone zależności.

Zespół drzew, choć zbliżony do Gradient Boosted Trees w strukturze, wykazuje najwyższy średni błąd względny wynoszący 0,0623. Maksymalny błąd (0,0827) i mediana (0,0658) również są większe niż w pozostałych modelach. Rozkład błędów charakteryzuje się jednak względnie wąskim zakresem (odchylenie standardowe 0,0123), co sugeruje stabilność wyników. Skośność (-0,3684) i kurtoza (-1,0544) wskazują na pewną asymetrię i spłaszczenie rozkładu błędów.

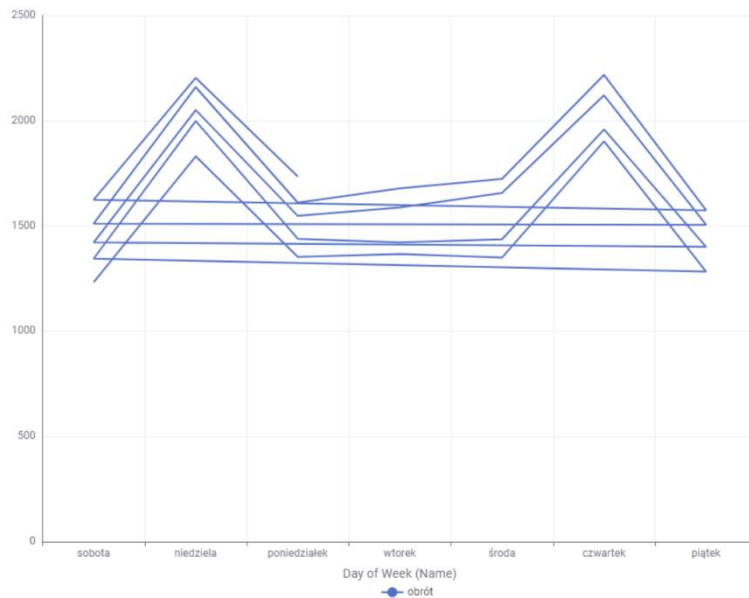
Podsumowując, pojedyncze drzewo jest najprostsze i ma najniższy błąd, Gradient Boosted Trees oferują najlepszy kompromis między dokładnością a stabilnością, natomiast zespół drzew wykazuje największy błąd, choć pozostaje stabilny w rozkładzie błędów.

Model liniowy

W kolejnym kroku opracowano model liniowy.

Etapy budowy modelu:

1. Wczytanie danych za pomocą węzła Excel Reader.
2. Wyodrębnienie z danych z kolumny data informacji o roku, miesiącu i dniu za pomocą węzła Date&Time Extractor.
3. Przesunięcie obrotu o jeden dzień do przodu za pomocą węzła Lag Column w celu udostępnienia możliwości porównywania dzisiejszego obrotu z dniem poprzednim.
4. Utworzenie nowej kolumny informującej o różnicy w obrocie dobowym za pomocą węzła Math Formula.
5. Wyświetlenie tylko pierwszego miesiąca (w celu ułatwienia interpretacji) oraz pokazanie obrotu w tym miesiącu za pomocą węzłów Row Filter i Line Plot.
6. Zauważenie wyraźnych zmian w obrocie w niedziele i w czwartki (rysunek poniżej) oraz przyjęcie odpowiednich wartości, aby wygładzić obrót dobowy.



Z kolumny \$przyrost dobowy\$ odczytano, że obrót w niedziele powinien być pomniejszony o 550, w czwartki powinien być pomniejszony o 500, a w piątki powiększony o 100.

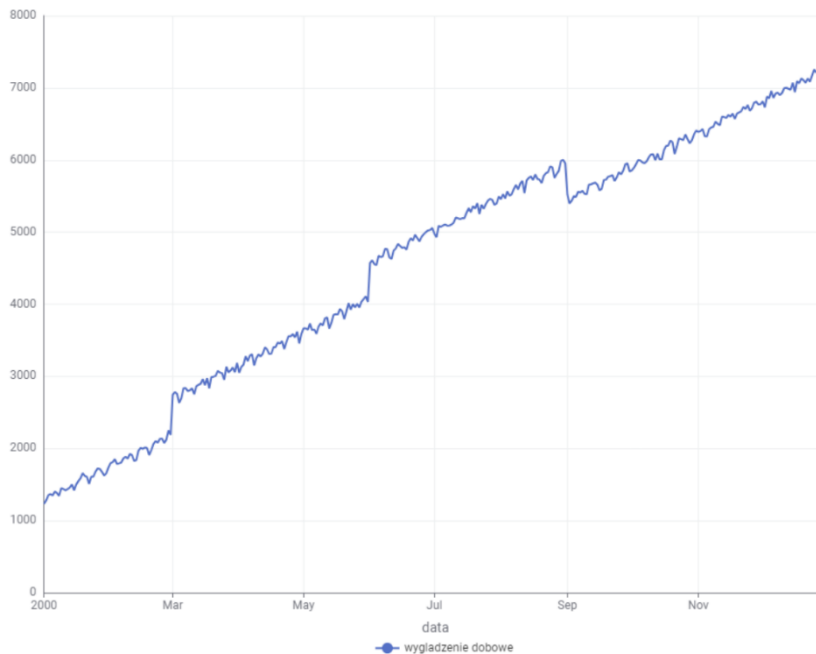
7. Dodanie zmiennej \$wahanie dzienne\$ za pomocą węzła Math Formula, przyjmującej wartość -550 w niedziele, -500 w czwartki i 100 w piątki, a w pozostałych dniach tygodnia 0.

```
if($Day of Week (Number)$==7,-550,
if($Day of Week (Number)$==4,-500,
if($Day of Week (Number)$==5,100,0)))
```

8. Dodanie zmiennej \$wygladzenie dobowe\$ za pomocą węzła Math Formula jako \$obróć\$ + \$wahanie dzienne\$.

$\$obróć\$ + \$wahanie\ dzienne\$$

9. Wyświetlenie tylko pierwszego roku (w celu ułatwienia interpretacji), pokazanie obrotu w tym roku za pomocą węzłów Row Filter i Line Plot. Zauważono wówczas wyraźne zmiany w obrocie przez pierwsze dwa miesiące oraz w okresie od czerwca do sierpnia (rysunek poniżej).



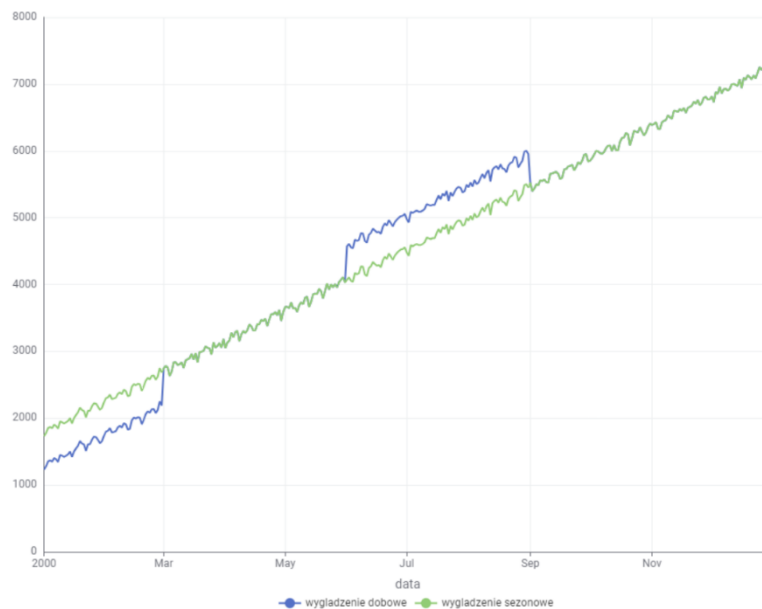
10. Dodanie zmiennej \$wahanie sezonowe\$ za pomocą węzła Math Formula, przyjmującej wartość -550 w styczniu i lutym oraz 500 w czerwcu, lipcu i sierpniu, a w pozostałych miesiącach 0. Wartości -550 i 500 zostały wywnioskowane z rysunku powyżej.

```
if($Month (Number)$==1 || $Month (Number)$==2, -500,
if($Month (Number)$==6 || $Month (Number)$==7 ||
$Month (Number)$==8,500,0))
```

11. Dodanie zmiennej \$wygladzenie sezonowe\$ za pomocą węzła Math Formula.

`$wygladzenie dobowe$-$wahanie sezonowe$`

Wygladzony obrót wygląda następująco (dane wyświetlone dla pierwszego roku)



12. Wyświetlenie informacji o obrocie, liczbie konkurencji oraz wygładzeniu sezonowym (za pomocą węzła Column Filter), obliczenie różnicy między obrotem a wygładzeniem sezonowym (za pomocą węzła Math Formula), a następnie wyświetlenie wszystkich informacji w celu sprawdzenia czy istnieje zależność obrotu od liczby konkurencji.

<input type="checkbox"/>	#	RowID	liczba konkurencji <i>Number (integer)</i>	<input type="checkbox"/>	obrót <i>Number (integer)</i>	<input type="checkbox"/>	wygładzenie sezonowe <i>Number (double)</i>	<input type="checkbox"/>	roznica <i>Number (double)</i>
<input type="checkbox"/>	1	Row0	0		1233		1,733		500
<input type="checkbox"/>	2	Row1	0		1831		1,781		50
<input type="checkbox"/>	3	Row2	0		1353		1,853		500
<input type="checkbox"/>	4	Row3	0		1367		1,867		500
<input type="checkbox"/>	5	Row4	0		1350		1,850		500
<input type="checkbox"/>	6	Row5	0		1903		1,903		0
<input type="checkbox"/>	7	Row6	0		1283		1,883		600
<input type="checkbox"/>	8	Row7	0		1345		1,845		500
<input type="checkbox"/>	9	Row8	0		1999		1,949		50
<input type="checkbox"/>	10	Row9	0		1439		1,939		500
<input type="checkbox"/>	11	Row...	0		1422		1,922		500
<input type="checkbox"/>	12	Row...	0		1437		1,937		500
<input type="checkbox"/>	13	Row...	0		1959		1,959		0
<input type="checkbox"/>	14	Row...	0		1401		2,001		600
<input type="checkbox"/>	15	Row...	0		1422		1,922		500
<input type="checkbox"/>	16	Row...	0		2051		2,001		50
<input type="checkbox"/>	17	Row...	0		1548		2,048		500
<input type="checkbox"/>	18	Row...	0		1588		2,088		500
<input type="checkbox"/>	19	Row...	0		1657		2,157		500
<input type="checkbox"/>	20	Row...	0		2121		2,121		0
<input type="checkbox"/>	21	Row...	0		1506		2,106		600
<input type="checkbox"/>	22	Row...	0		1511		2,011		500
<input type="checkbox"/>	23	Row...	0		2160		2,110		50
<input type="checkbox"/>	24	Row...	0		1611		2,111		500
<input type="checkbox"/>	25	Row...	0		1679		2,179		500
<input type="checkbox"/>	26	Row...	0		1724		2,224		500

13. Zauważenie możliwości wygładzenia danych oraz dodanie zmiennej \$wpływ konkurencji\$ (za pomocą węzła Math Formula).

`-500*($liczba konkurencji$+1)`

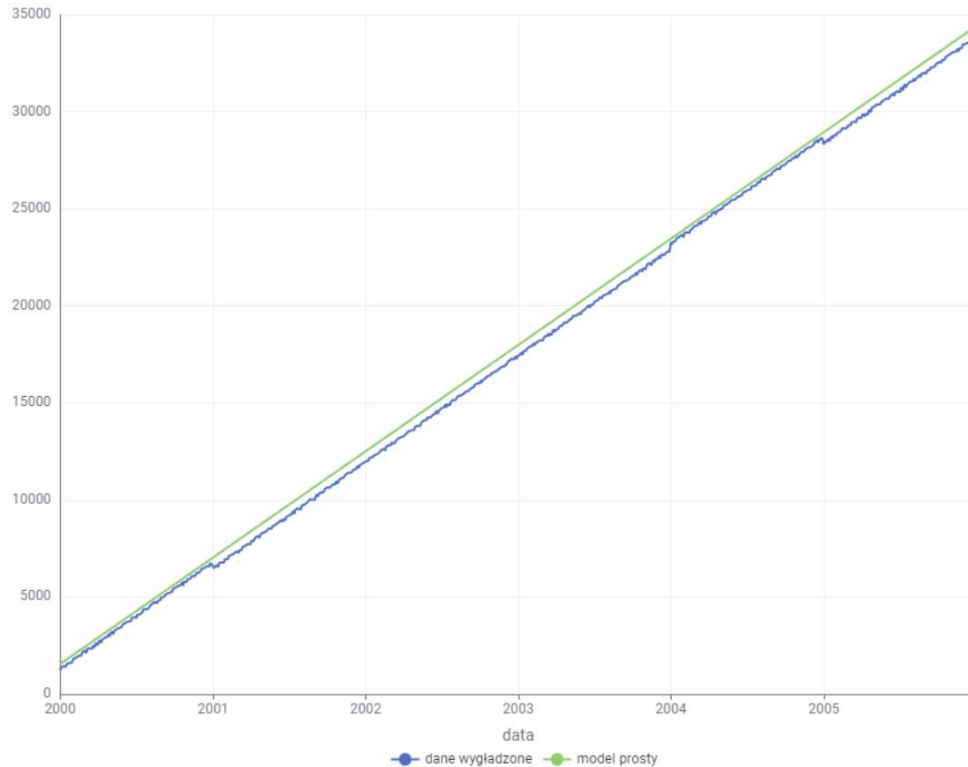
14. Dodanie zmiennej 'dane wygładzone' (za pomocą węzła Math Formula).

`$wygladzenie sezonowe$+$wpływ konkurencji$`

15. Utworzenie prostego modelu liniowego ze zmienną niezależną \$t\$ poprzez odczytanie współczynników z węzła Linear Regression Learner oraz dodanie za pomocą węzła Math Formula zmiennej \$model prosty\$.

`1536+15*t`

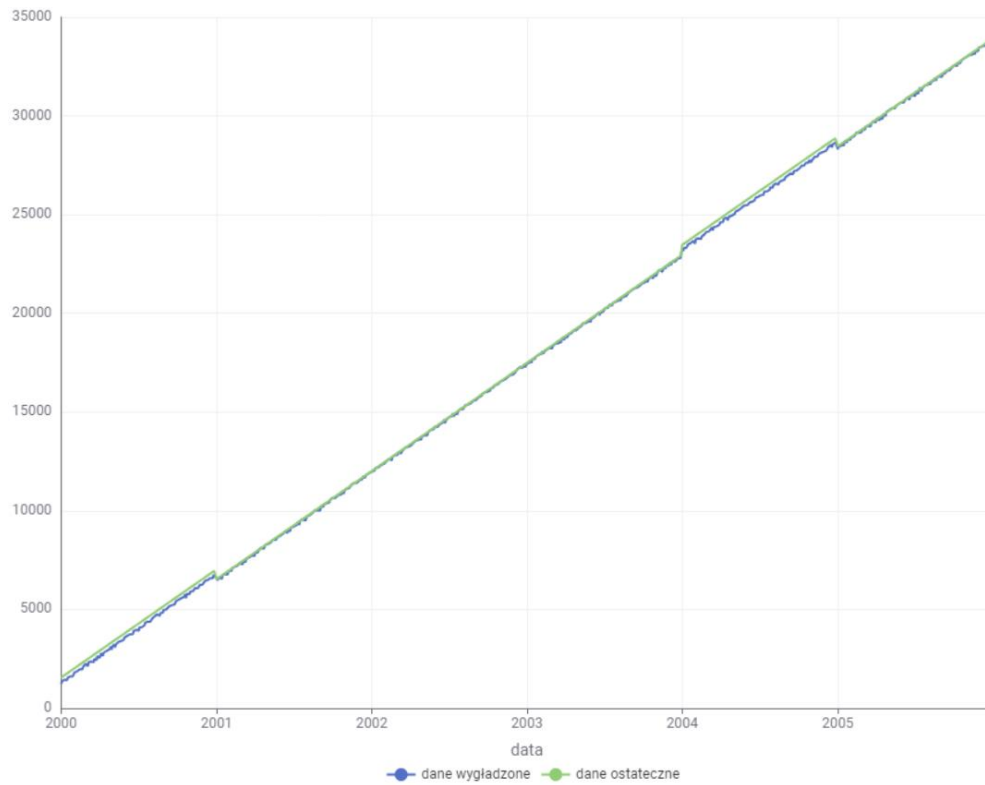
16. Wyświetlenie obrotu oraz danych wygładzonych dla pierwszych 5 lat (tylko 5 w celu ułatwienia interpretacji) w celu zauważenia, czy istnieje możliwość wygładzenia danych dla poszczególnych lat oraz zauważenie, że dla lat przestępnych obrót jest podwyższony w stosunku do pozostałych lat.



17. Utworzenie nowej zmiennej `$dane ostateczne$` za pomocą Math Formula, uwzględniającej skoki w latach przestępnych.

```
if($Year%4 != 0, $model prosty$-500, $model prosty$)
```

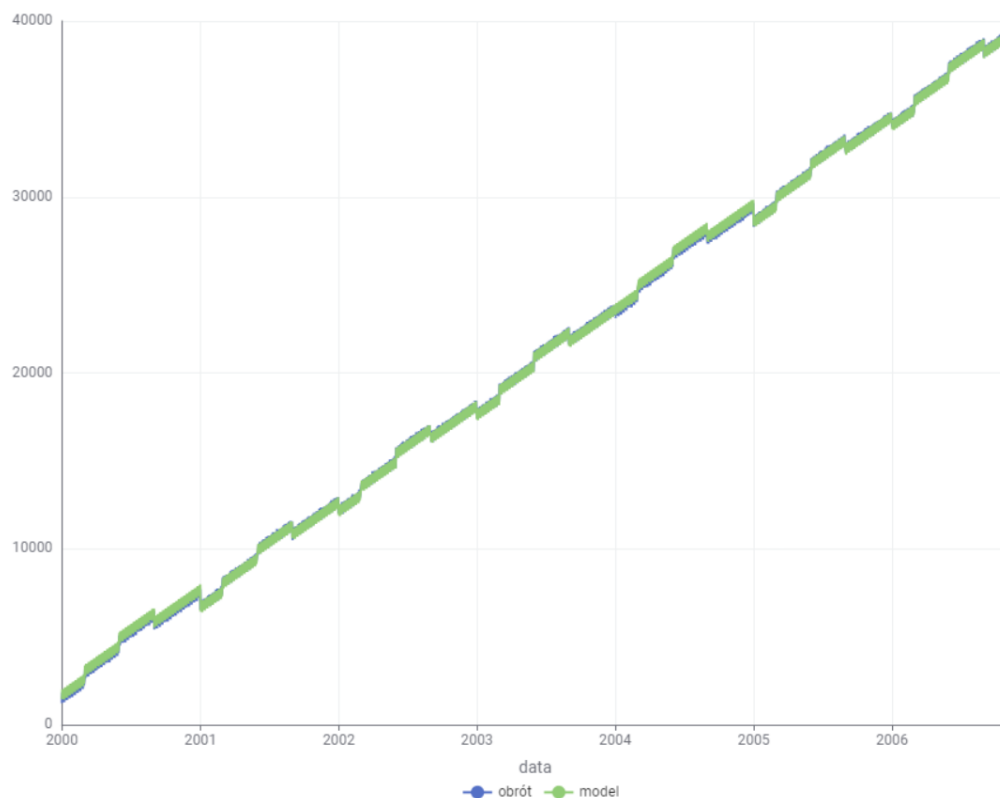
Wygładzone dane wyglądają w następujący sposób.




18. Utworzenie modelu liniowego poprzez dodanie za pomocą węzła Math Formula zmiennej \$model\$.

```
$dane ostateczne$+$wpływ konkurencji$+$wahanie sezonowe$  
-$wahanie dzienne$+900
```

Wartość 900 została dobrana metodą prób i błędów w oparciu o wykres.
Ostateczna postać modelu na wykresie pokrywa się z pierwotnym obrotem.



19. Obliczenie błędu względnego oraz pokazanie dla niego statystyk za pomocą węzłów Math Formula, Column Filter i Statistics. Statystyki błędu przedstawiono na rysunku poniżej.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
błąd względny	0.0	0,0034	0,0013	0,1768	0,0099	8,2286	88,1217	1	0	0	

Model charakteryzuje się bardzo niskim błędem względnym, co świadczy o jego wysokiej dokładności.

- Minimalny błąd wynosi 0, co oznacza, że dla niektórych obserwacji model przewidział wyniki idealnie.
- Średni błąd względny wynosi zaledwie 0,0034, co wskazuje, że przewidywania modelu są w większości przypadków bardzo bliskie rzeczywistym wartościom.
- Mediana błędu, wynosząca 0,0013, sugeruje, że ponad połowa obserwacji ma jeszcze mniejsze błędy.
- Największy błąd względny w modelu wynosi 0,1768, co oznacza, że w najgorszym przypadku przewidywania mogą być obarczone błędem na poziomie 17,7%. Odchylenie standardowe wynosi 0,0099, co wskazuje na niewielkie zróżnicowanie błędów w danych, a więc model jest stabilny w swoich przewidywaniach.

- Skośność rozkładu błędów, wynosząca 8,2286, pokazuje, że zdecydowana większość wartości błędów jest bardzo mała, z nielicznymi przypadkami większych odchyleń.
- Kurtosis na poziomie 88,1217 dodatkowo wskazuje na dużą koncentrację wartości w pobliżu średniej, co potwierdza, że większość błędów jest wyjątkowo niska.

Porównując ten model z wcześniejszymi trzema modelami uczenia maszynowego, model liniowy osiąga najniższy średni błąd względny, co świadczy o jego wysokiej dokładności. Pojedyncze drzewo decyzyjne i Gradient Boosted generują nieco wyższe średnie błędy, co wskazuje, że w tym przypadku modele te nie przewyższają podejścia liniowego. Zespół drzew wypadł najgłupszy, z najwyższym średnim błędem względnym, co sugeruje, że nie jest optymalnym rozwiązaniem dla tych danych. Wyniki wskazują, że model liniowy lepiej radzi sobie z analizowanymi danymi niż bardziej złożone algorytmy.