METODY SZTUCZNEJ INTELIGENCJI W INŻYNIERII OPROGRAMOWANIA

Opracował: dr hab. inż. Paweł Piotrowski, wersja 30.10.2023

Ćwiczenie 6 h– Prognozowanie krótkoterminowe kursu dolara oraz analiza biznesowa potencjalnych korzyści z prognoz
Punkty 1-2 (2h), 3,4,5 (2h), 6,7,8 (2h)

Środowisko wykonania ćwiczenia: arkusz kalkulacyjny, program Statistica (alternatywnie środowisko Python dla chętnych)

Celem ćwiczenia jest wykonanie analizy danych (dobór zmiennych do modeli), prognozy z horyzontem t+1 oraz t+7 kursu dolara z wykorzystaniem wybranych technik uczenia maszynowego (problem regresyjny), a następnie weryfikacja czy wykorzystanie w praktyce prognoz ma sens biznesowy (decyzja o zakupie towaru bez i z wykorzystaniem prognoz t+1 oraz t+7)

---

1. Analiza danych

W pliku „prognozy_kurs_dolara" w zakładce „dane pierwotne" zgromadzono dane jako szereg czasowy, kolumna C to losowo przypisane liczby 1,2,3 do kolejnych wierszy z zachowaniem proporcji ilościowych (dla sieci MLP, która wymaga danych walidacyjnych- właściwy moment zakończenia nauki), kolumna D jest identyczna z C ale wszystkie wiersze z liczbą 2 zamieniono na liczbę 1 powiększając pulę treningową (dla technik innych niż MLP nie wydzielono danych walidacyjnych – stanowią one dane treningowe).

Wykonujemy wykres szeregu czasowego USD t w celu wizualizacji zagadnienia

W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652
W arkuszu excel zaznaczamy zakres E1:BJ3653 czyli z nagłówkami wybrane dane z zakładki "dane_pierwote" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

Wybieramy->Statystyka ->statystyki podstawowe->statystyki opisowe
Zmienne-> wybieramy USD t
Wybierając kolejne zakładki możemy wykonać wybrane analizy statystyczne szeregu czasowego USD t

Wybór zmiennych wg wartości współczynnika korelacji liniowej Pearsona

Wybieramy->Statystyka ->statystyki podstawowe-> macierze korelacji
Wybieramy->jedna lista zmiennych->
Analizę korelacji wykonujemy osobno dla problemu horyzont t+7 (wyjście) oraz t+1 (wyjście)

Dla przypadku prognozy t+1 wpisujemy 1 3-58 w polu wyboru lub ręcznie wybieramy zmienne z przyciśniętym CTRL

W zakładce podstawowe wybieramy -> korelacje
Notujemy wyniki kopiujemy (wybierz wszystko, kopiuj z nagłówkami) do arkusza kalkulacyjnego
Dokonujemy wyboru zmiennych wejściowych analizując współczynniki korelacji zmiennej wyjściowej do potencjalnych zmiennych wejściowych (dane istotne statystycznie zaznaczone są na czerwono)

Wybierając -> macierz wykresów rozrzutu możemy zbadać jaki rodzaj zależności występuje pomiędzy zmienną wyjściową oraz wybraną potencjalną zmienną wejściową np. USD t+1 do euro t itp.

Wybór zmiennych jako problem regresji wielokrotnej liniowej (statystyka F Fischera Snedecora – eliminacja zmiennych z równania regresji

Z menu górnego wybieramy „data mining" , wybieramy ->dobór zmiennych -> dobór zmiennych
Zależna ilościowa = zmienna wyjściowa, predyktory ilościowe = potencjalne zmienne wejściowe 3-58
Liczba klas dla predyktorów ilościowych np. 20
Wybieramy pokaż 56 najlepszych predyktorów (komentarz: Pokaż k najlepszych predyktorów. Ta opcja oznacza wyświetlanie k najlepszych predyktorów. W problemach typu regresyjnego (dla ilościowych zmiennych zależnych), jest to k predyktorów o najwyższej wartości statystyki F)
Klikamy w podsumowanie najlepsze predyktory
Notujemy w arkuszu kalkulacyjnych wyniki – sugestie – wybór zmiennych wejściowych

Wybór zmiennych jako problem regresji wielokrotnej liniowej (Statystyka R2 - współczynnik determinacji - jest to kwadrat współczynnika korelacji liniowej Pearsona pomiędzy dwiema zmiennymi, wyraża on wielkość wariancji wspólnej dwóch zmiennych)

Z menu górnego wybieramy „data mining" , wybieramy ->dobór zmiennych -> dobór i kategoryzacja predyktorów
Zależna ilościowa = zmienna wyjściowa, predyktory ilościowe = potencjalne zmienne wejściowe 3-58
Notujemy w arkuszu kalkulacyjnych wyniki – sugestie – wybór zmiennych wejściowych

Uwaga: w przypadku części technik uczenia maszynowego np. lasy losowe możemy ocenić jakość zmiennych wejściowych – umożliwia to algorytm

---

2. Prognozy z wykorzystaniem sieci neuronowej MLP dla horyzontu t+1 oraz horyzontu t+7

W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652

W arkuszu excel zaznaczamy zakres B1:BH3653 czyli z nagłówkami wybrane dane z zakładki "dane_siec_MLP" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

menu -> Data Mining -> Sieci neuronowe
wybieramy -> nowa analiza -> regresja

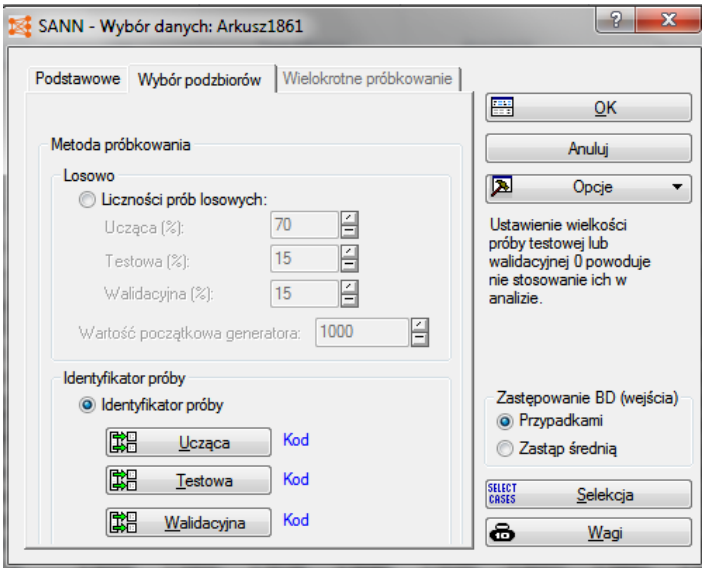
okno SANN wybór danych
zakładka "podstawowe"
wybieramy ->projekt sieci użytkownika
klikamy w pole "zmienne"
wybieramy "wyjscia ilościowe" – USD t+1 lub USD T+7 w zależności od horyzontu prognozy
wybieramy "wejścia ilościowe" - tutaj wskazujemy (trzymając naciśnięty klawisz CTRL) wybrane przez nas zmienne objaśniające (sugerowałbym najpierw komplet zmiennych)

klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby"
klikamy w pole "ucząca" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną "kod_danych losowy_TRE_WAL_TEST", w polu "kod próby uczącej" wpisujemy "1", klikamy ok.

klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby"
klikamy w pole "testowa" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną "kod_danych losowy_TRE_WAL_TEST", w polu "kod próby testowej" wpisujemy "2", klikamy ok.

klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby"
klikamy w pole "walidacyjna" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną "kod_danych losowy_TRE_WAL_TEST", w polu "kod próby walidacyjnej" wpisujemy "3", klikamy ok.

UWAGA: w konwencji statistica próba niewidoczna dla sieci neuronowej to walidacyjna czyli dla nas to zakres testowy, najważniejszy do analizy



zamykamy okno "SANN wybór danych klikając pole OK.

pojawia się zakładka "SANN projekt użytkownika"

w zakładce "podstawowe"
wybieramy -> typ sieci -> perceptron wielowarstwowy
wybieramy funkcję aktywacji dla neuronów ukrytych oraz neuronu wyjściowego
wybieramy liczbę sieci oraz liczbę neuronów w warstwie ukrytej

w zakładce "perceptron"
wybieramy algorytm uczenia (BFGS), liczbę epok oraz rodzaj inicjalizacji sieci

w zakładce "wykres uczenia'
wybieramy - wykres uczenia dla błędów w próbie uczącej oraz testowej.

w oknie głównym "SANN projekt użytkownika" klikamy w przycisk "uczenie"
aby zobaczyć liczbę epok uczących należy rozszerzyć kolumnę "algorytm"
klikamy w "wybór sieci" i wskazujemy najlepszą sieć - z najmniejszym błędem w zakresie "jakość uczenia" - ta miara błędu to po prostu współczynnik korelacji liniowej Pearsona
klikając w pole "podsumowanie" zobaczymy wartości miar błędów dla zakresów uczenia i testowania (błąd SOS to średnia z sum kwadratów odchyleń dla próby uczącej i testowej)

W prawym dolnym rogu okna głównego - opcja "próby" zaznaczamy pozycję "walidacja" – jest to nasza próba testowa (model nie widział tych danych)
Następnie klikając w pole "globalna analiza wrażliwości" - uzyskujemy informacje o **ważności poszczególnych zmiennych objaśniających** w modelu prognostycznym na zakresie testowym – opis analizy wrażliwości w "dymku"

Fragment help programu Statistica

Globalna analiza wrażliwości. Globalna analiza wrażliwości daje pojęcie o tym, jak ważne są poszczególne zmienne wejściowe sieci. Wykonanie analizy wrażliwości polega na sprawdzeniu jak zachowuje się błąd sieci w przypadku gdy coś złego dzieje się ze zmiennymi niezależnymi. Konkretnie, po kolei dla każdej zmiennej wejściowej jej wartości zamieniane są na średnią (ze zbioru uczącego). Tak więc zmienna przestaje wnosić jakąkolwiek informację. Po podaniu tak zmodyfikowanych danych na wejście sieci sprawdza się końcowy błąd predykcji. Błąd ten może

poważnie wzrosnąć, albo wzrosnąć nieznacznie lub wcale. Oznacza to, że sieć jest albo bardzo wrażliwa na daną zmienną wejściową, albo też sieci na tej zmiennej zupełnie nie zależy.
W arkuszu, dla każdej sieci podany jest iloraz wskazujący przyrost błędu przy usunięciu danej zmiennej wejściowej. Jeżeli wartość jest 1 lub mniejsza to sieć działa lepiej bez danej zmiennej - znak, ze należy ją usunąć na stałe.
Jednak pamiętać trzeba, że analiza dotyczy konkretnej sieci. Tymczasem zmienne bywają na różne sposoby powiązane, skorelowane i wykazują redundancje. Dlatego różne sieci mogą "wybrać" jako ważne różne zmienne. Dopiero wykonanie analizy wrażliwości dla wielu modeli i powtarzalność wyników powinny być podstawą do wyciągania praktycznych wniosków na temat zmiennych.

Ewentualnie korygujemy później model wybierając nowy, lepszy zestaw zmiennych objaśniających wejściowych.

w prawym dolnym rogu okna głównego - opcja "próby" zaznaczamy pozycję "walidacja" – jest to nasza próba testowa (model nie widział tych danych)

klikamy w przycisk "predykcja"
kolumna „WY USD....." - wyjscie to prognozy dla naszego zakresu testowego-wklejamy ten blok danych do excel -zakładka "wyniki_siec_MLP" - kolumna E lub F w zależności od horyzontu prognozy
W kalkulatorze błędów pojawią się obliczone błędy MAPE% oraz RMSE – kopiujemy wyniki do tabeli z wariantami oraz opisujemy w tabeli parametry danego wariantu. Kolumnę prognoz wklejamy do archiwum wyników.

kolejne kroki: szukamy właściwych parametrów sieci dla których wartości miar błędów w zakresie danych testowych będą najmniejsze-można wykorzystać narzędzie "więcej (projekt automatyczny)
poszukiwanie właściwego modelu to również manipulowanie doborem zmiennych

---

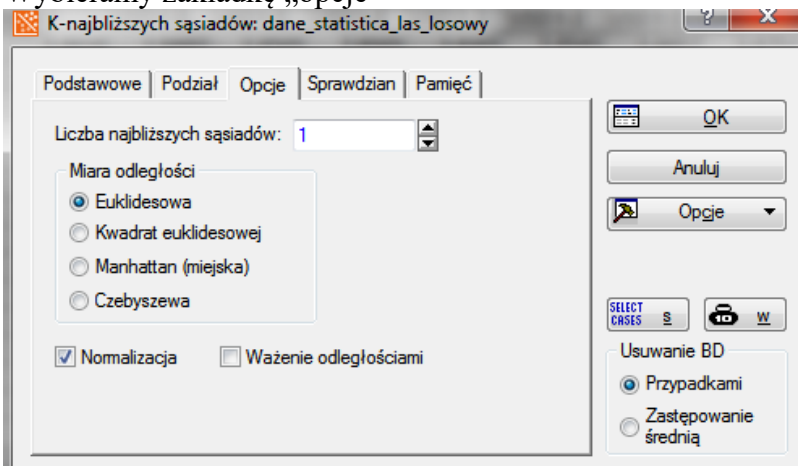
3. Prognozy z wykorzystaniem techniki uczenia maszynowego - KNN regresyjny dla horyzontu t+1 oraz horyzontu t+7

W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652
W arkuszu excel zaznaczamy zakres B1:BH3653 czyli z nagłówkami wybrane dane z zakładki "dane_uczenie_maszynowe" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

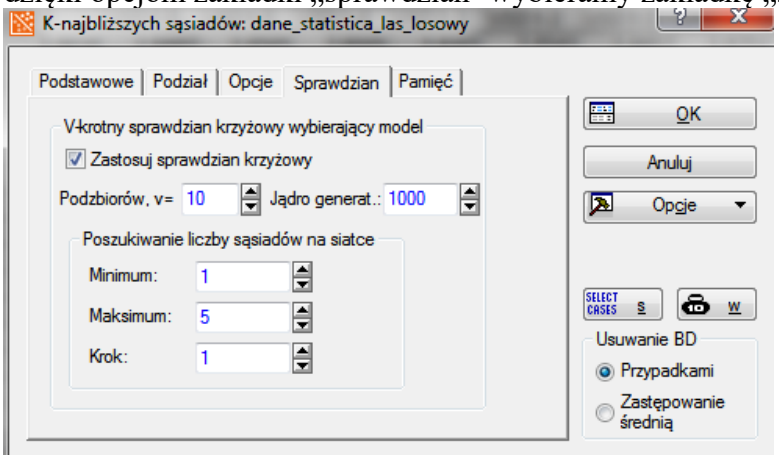
menu -> Data Mining -> Inne metody uczenia maszynowego->wybieramy pozycję „K-najbliższych sąsiadów"
W zakładce „podstawowe" wybieramy przycisk „zmienne" wskazujemy wyjście modelu (zależne ilościowe) oraz zestaw wejść do modelu (predyktory ilościowe)
trzymając klawisz CTR można wybrać predyktory ilościowe lub wpisać liczbowo w dolnym oknie.
W zakładce „podział" klikamy „według zmiennej (próby)" stan – włącz, klikamy „zmienna identyfikująca próby" wybieramy pozycję pierwszą z listy – kod_danych_losowy_TRE_TEST, klikamy OK.

następnie podajemy przy polu „kod próby do analizy” liczbę 1 (oznacza to, że próba ucząca to dane z tą sygnaturą), klikamy OK.
Wybieramy zakładkę „opcje”



Rys. Okno wyboru parametrów

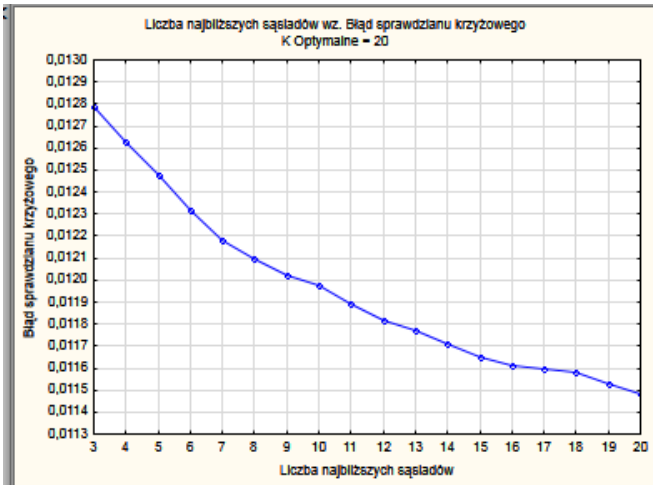
Wybieramy miarę odległości, liczba najbliższych sąsiadów będzie ustalana automatycznie dzięki opcjom zakładki „sprawdzian- wybieramy zakładkę „sprawdzian”



Rys Okno „sprawdzian.

Klikamy w „Zastosuj sprawdzian krzyżowy”
Wybieramy liczbę podzbiorów, minimum oraz maksimum liczby sąsiadów
Klikamy OK. uruchamiając algorytm
Uwaga: szczegółowy opis działania algorytmu można przeczytać klikając w „?” w prawym górnym rogu.

Po zakończeniu procesu w zakładce „podstawowe” wybieramy w prawym dolnym rogu „próba”->testowa, klikamy w ikonę „przewidywania na dole okna po lewej stronie. Kopiujemy kolumnę „przewidywane” czyli prognozy dla zakresu testowego do kolumny E lub F w zależności od horyzontu prognozy w zakładce „KNN” arkusza kalkulacyjnego.
Klikamy w przycisk „sprawdzian krzyżowy” i obserwujemy zależność błędu od liczby najbliższych sąsiadów co ułatwia proces szukania optymalnej wartości liczby najbliższych sąsiadów.



Rys, Okno wyników dla różnej liczby najbliższych sąsiadów

Szukamy najlepszego wyniku manipulując miarą odległości, zakresem min-maks liczby najbliższych sąsiadów oraz liczbą v podzbiorów. Zmieniać możemy też listę zmiennych wejściowych.
Uwaga: program zapamiętuje wynik z momentu zatrzymania poszukiwań więc należy przy odpytywaniu wybrać właściwą liczbę k

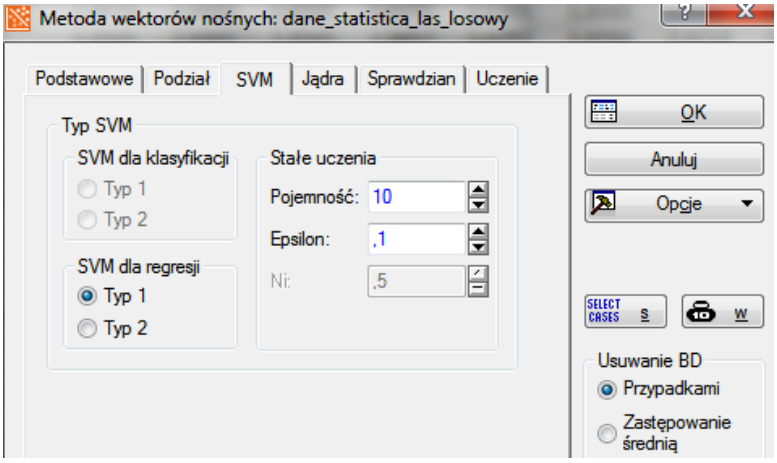
---

4. Prognozy z wykorzystaniem techniki uczenia maszynowego SVM typu SVR (regresyjna SVM) dla horyzontu t+1 oraz horyzontu t+7

W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652
W arkuszu excel zaznaczamy zakres B1:BH3653 czyli z nagłówkami wybrane dane z zakładki "dane_uczenie_maszynowe" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

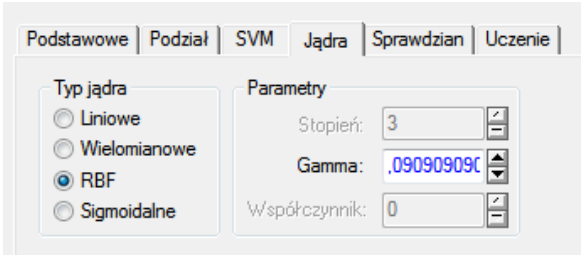
menu -> Data Mining -> Inne metody uczenia maszynowego->wybieramy pozycję „metoda wektorów nośnych”
W zakładce „podstawowe” wybieramy przycisk „zmienne” wskazujemy wyjście modelu (zależne ilościowe) oraz zestaw wejść do modelu (predyktory ilościowe)
trzymając klawisz CTR można wybrać predyktory ilościowe lub wpisać liczbowo w dolnym oknie.
W zakładce „podział” zaznaczamy pozycję „podziel dane na próbę uczącą i testową” ,następnie wybieramy „według zmiennej klimay w „próby” stan – włącz, klikamy „zmienna identyfikująca próby” wybieramy pozycję pierwszą z listy – kod_danych_losowy_TRE_TEST, klikamy OK.
następnie podajemy przy polu „kod próby do analizy” liczbę 1 (oznacza to, że próba ucząca to dane z tą sygnaturą), klikamy OK.
Klikamy w zakładkę „SVM” i podajemy parametry



Rys. Okno wyboru parametrów SVM

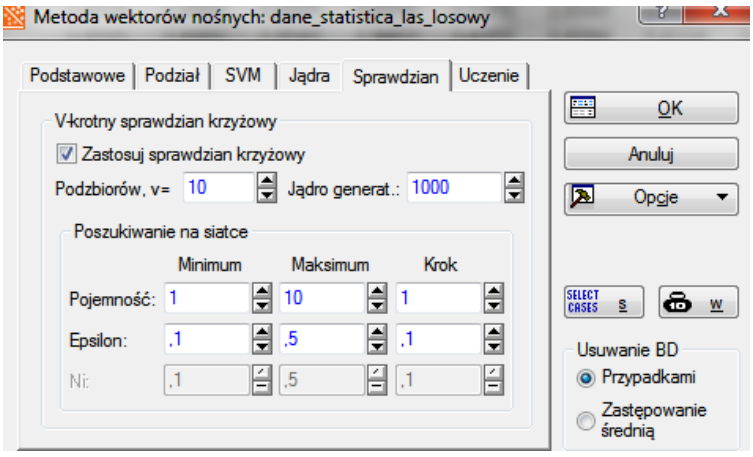
Stałe uczenia. Zależnie od wybranego typu modelu, ustalamy dostępne parametry uczenia modelu. Optymalne wartości tych parametrów nie są znane a priori, jednak można mieć jakąś ich ocenę. Można użyć metody sprawdzianu krzyżowego do otrzymania wartości stałych uczenia. Są trzy parametry uczenia:
Pojemność. Ten parametr ma zastosowanie do obu modeli regresyjnych.
Epsilon. Ten parametr ma zastosowanie jedynie do Modelu SVM typu 1, regresyjnego.

Wybieramy zakładkę „Jądra” i wybieramy typ jądra.



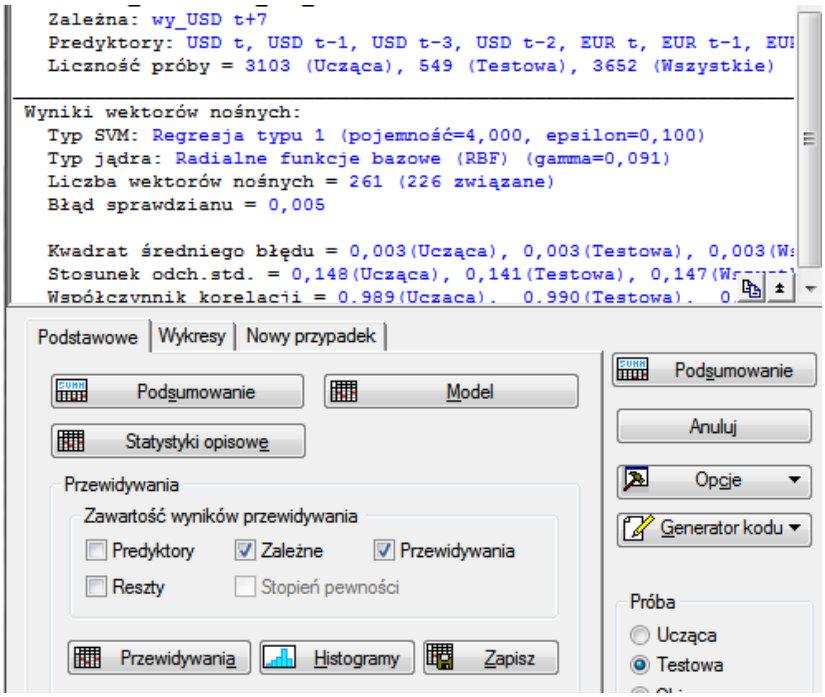
Rys. Okno wyboru typu jądra

Wybieramy zakładkę „sprawdzian” i ustalamy zakresy testowanych parametrów (pojemność i epsilon) oraz liczbę podzbiorów v



Rys. Okno wyboru zakresów testowania parametrów

W zakładce „uczenie” wybieramy maksymalną liczbę iteracji (korzystniej 2-3 tysiące podać zamiast sugerowanej)
Uwaga: być może lepsze wyniki uzyskuje się gdy wybierze się normalizację i zawężanie – zweryfikować to.
Klikamy OK aby rozpocząć działanie algorytmu. W oknie wyników możemy odczytać optymalną wartość pojemności oraz epsilon. Możemy w kolejnych próbach poszerzyć zakres jeśli parametr optymalny był na progu czyli wartości maksymalnej.



Rys Okno wyników

Po zakończeniu procesu w zakładce „podstawowe” zaznaczamy próba->testowa, następnie klikamy w przycisk „przewidywania”

Kopiujemy kolumnę „przewidywane” czyli prognozy dla zakresu testowego do kolumny E lub F w zależności od horyzontu prognozy w zakładce „SVR” arkusza kalkulacyjnego.
Klikamy w przycisk „sprawdzian krzyżowy” i obserwujemy zależność błędu od liczby najbliższych sąsiadów co ułatwia proces szukania optymalnej wartości liczby najbliższych sąsiadów.

Szukamy najlepszego wyniku manipulując parametrami. Zmieniać możemy też listę zmiennych wejściowych.

---

5. Prognozy z wykorzystaniem techniki uczenia maszynowego – lasy losowe dla horyzontu t+1 oraz horyzontu t+7

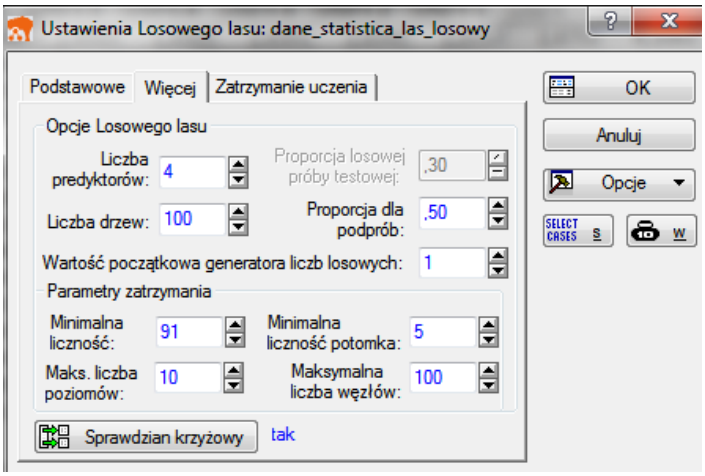
W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652
W arkuszu excel zaznaczamy zakres B1:BH3653 czyli z nagłówkami wybrane dane z zakładki "dane_uczenie_maszynowe" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

menu -> Data Mining -> pod napisem „podstawowe” (górne menu) wybieramy czerwoną ikonę z napisem „lasy losowe” (środkowa ikona w 2 rzędzie czerwonych ikon)

w oknie wybieramy pozycję -> zadanie regresyjne, klikamy ok.
w zakładce „podstawowe” wybieramy przycisk „zmienne” wskazujemy wyjście modelu (zmienna zależna) oraz zestaw potencjalnych wejść do wyboru przez algorytm lasu losowego w analizie (predyktory ilościowe)
trzymając klawisz CTR można wybrać predyktory ilościowe lub wpisać liczbowo w dolnym oknie.

W zakładce „więcej” wybieramy
a) liczbę predyktorów n (będzie taka sama) dla każdego z drzew (przyjmując, że wybraliśmy np. k predyktorów ilościowych w zakładce „podstawowe” to liczba predyktorów jest równa lub mniejsza niż k – te predyktory w liczbie n są dla każdego drzewa losowo wybierane z dostępnych k)
b) liczbę drzew
c) parametry zatrzymania (elementy regularyzacji)
Następnie klikamy w przycisk „sprawdzian krzyżowy” w celu wskazania próby uczącej (sygnatura 1) (pozostałe przypadki będą próbą testową) – klikamy w „stan włączony”, klikamy w „zmienna identyfikująca próby” – wybieramy pozycję pierwszą z listy – kod_danych_losowy_TRE_TEST, klikamy OK.,
następnie podajemy przy polu „kod próby do analizy” liczbę 1 (oznacza to, że próba ucząca to dane z tą sygnaturą), klikamy OK.



Rys. Okno parametrów lasu losowego

Uwagi:
Liczba predyktorów. W tym polu podajemy liczbę predyktorów losowo wybieranych dla każdego z drzew składających się na Losowy las. Domyślna liczba jest wyznaczana jako log2(“liczba predyktorów” + 1).
Liczba drzew. W tym polu podajemy maksymalną liczbę drzew składających się na Losowy las. Program tworzy zespoły złożone z rosnącej liczby drzew i oblicza błąd dla próby testowej i uczącej. Zmianę tych wskaźników wraz ze wzrostem liczby drzew składowych

obserwujemy w oknie Obliczanie. Jeśli włączymy Zaawansowane warunki zatrzymania na karcie Zatrzymanie uczenia, to program automatycznie znajdzie najlepszą liczbę drzew nieprzekraczającą wartości podanej w polu Liczba drzew. Po zakończeniu tworzenia modelu możemy również badać wpływ wielkości zespołu na trafność przewidywań, korzystając z opcji w oknie Wyniki Losowego lasu

Parametry zatrzymania – cztery parametry.
W tej grupie znajdują się ustawienia stosowane przy budowie poszczególnych drzew. Ustawienia te wpływają na wielkość drzew.

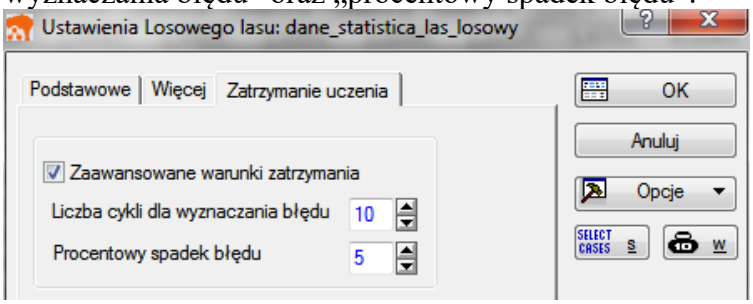
1. Minimalna liczność. Jednym ze sposobów ograniczenia wielkości drzewa jest ustalenie minimalnej liczby obiektów w węźle podlegającym podziałom. Właśnie ta liczba znajduje się w tym polu.

2. Minimalne n potomka. W wyniku podziału węzła uzyskujemy węzły - potomków. Ustawienie Minimalne n potomka określa najmniejszą dopuszczalną liczność tych węzłów (w odróżnieniu od parametru Minimalna liczność ograniczającego wielkość węzła podlegającego podziałowi). Określenie minimalnej liczności potomka jest pomocne w sytuacji, gdy drzewa składowe mają węzły końcowe o zbyt małej liczności (zwłaszcza po jednej stronie lub w obrębie jednej gałęzi drzewa). Takiemu problemowi zapobiega zwiększenie wartości parametru Minimalne n potomka.

3. Maks. n poziomów. Wartość w tym polu ogranicza liczbę poziomów drzew składowych. Przy każdym podziale program sprawdza ile poziomów jest od bieżącego węzła do węzła początkowego, a jeśli zostanie osiągnięta maksymalna liczba poziomów, to dalsze podziały nie będą wykonywane.

4. Maksymalna liczba węzłów. Ta wartość ogranicza liczbę węzłów tworzących drzewa składowe. Przed każdym podziałem sprawdzana jest całkowita liczba węzłów i jeśli jest ona równa Maksymalnej liczbie węzłów, to proces budowy drzewa jest zatrzymywany.

W zakładce „zatrzymanie uczenia" ewentualnie zmieniamy parametry „liczba cykli dla wyznaczania błędu" oraz „procentowy spadek błędu".



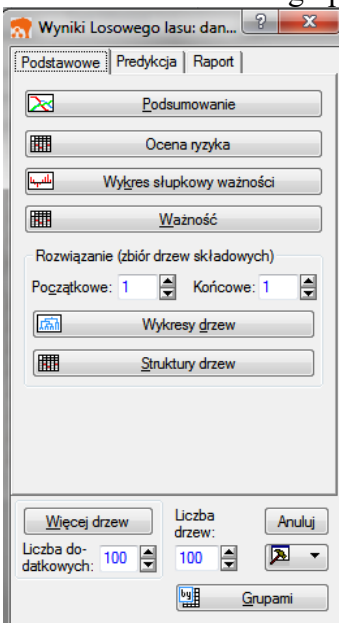
Rys. Okno parametrów zatrzymania uczenia

Uwagi:
Zaawansowane warunki zatrzymania. Zaznaczenie tego pola włącza sprawdzanie kryterium zatrzymania procesu tworzenia modelu przed osiągnięciem liczby drzew podanej na karcie Więcej. Wybranie tej opcji powoduje uaktywnienie pozostałych pól na karcie Zatrzymanie uczenia.

Liczba cykli dla wyznaczenia błędu. To pole zawiera liczbę cykli uczenia, dla których porównywany jest błąd. Błąd w próbie testowej i uczącej dla modeli Losowego lasu wykazuje przypadkowe wahania. Dlatego zazwyczaj nie jest dobrze zatrzymywać uczenie na podstawie zmiany błędu w pojedynczym cyklu. Program sprawdza, czy po wykonaniu podanej liczby cykli uczenia (innymi słowy dodaniu kolejnych drzew do zespołu) błąd zmniejszył się o wartość podaną w polu Procentowy spadek błędu. Jeśli po wykonaniu podanej liczby cykli spadek błędu jest mniejszy niż podana wartość, to uczenie zostaje przerwane.

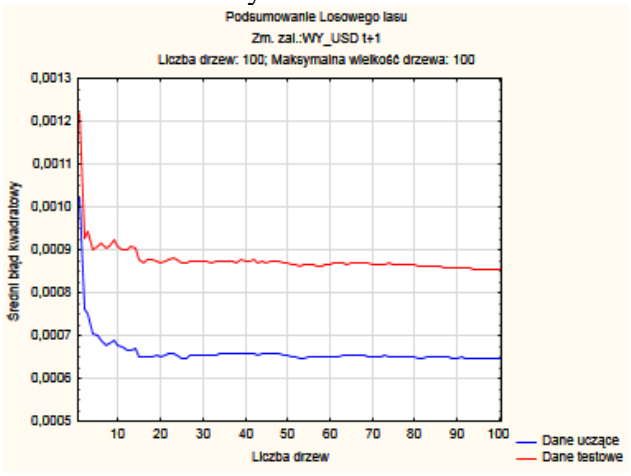
Procentowy spadek błędu. Tu podajemy w procentach minimalny spadek błędu wymagany do kontynuowania procesu uczenia modelu.

Po kliknięciu OK. algorytm lasu losowego rozpoczyna trening.
Po zakończeniu treningu pojawi się okno jak poniżej



Rys. Okno „Wyniki losowego lasu”

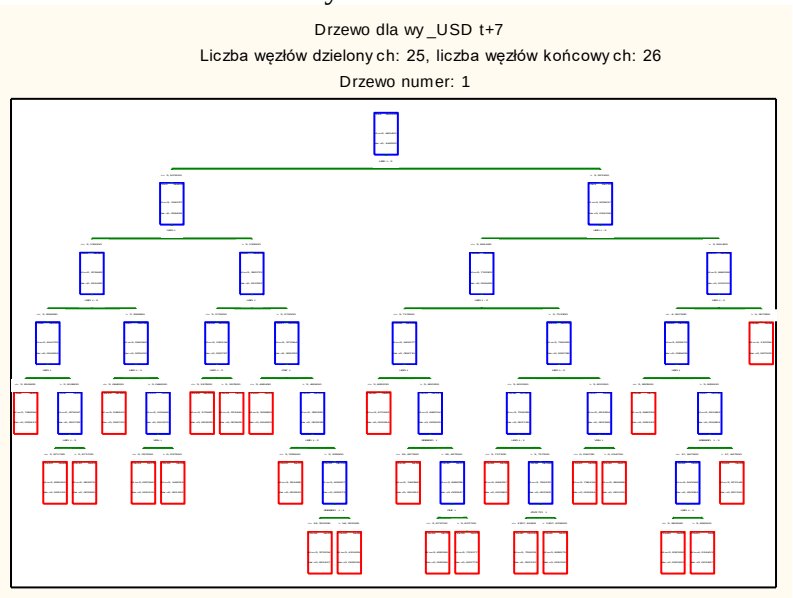
W zakładce „podstawowe” klikając w „podsumowanie” zobaczymy zmiany błędu w zależności od liczby drzew.



Rys Okno błędu w zależności od liczby drzew

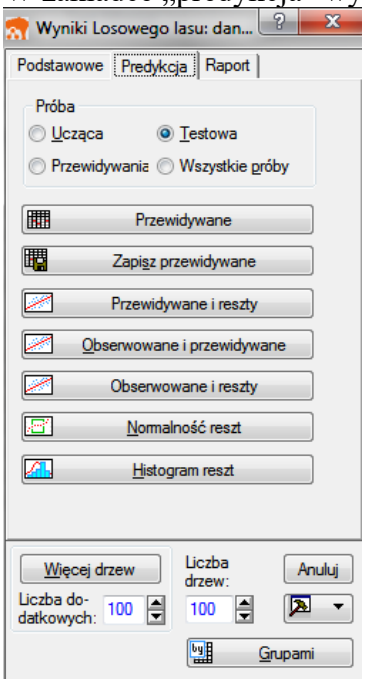
W zakładce „podstawowe” możemy zobaczyć cechy lasu losowego np. wykres słupkowy ważności **(ocena ważności poszczególnych zmiennych wejściowych)**, wykresy drzew,

struktury drzew, możemy również dodać kolejne lasy (może to poprawić wyniki) – więcej drzew-> liczba dodatkowych



Rys. Okno pojedynczego drzewa z n drzew

W zakładce „predykcja" podajemy właściwą liczbę drzew w pozycji „liczba drzew" czyli taką dla której błąd był najmniejszy – obserwacja wykresu „Okno błędu w zależności od liczby drzew"
W zakładce „predykcja" wybieramy ->próba testowa i klikamy w przycisk „Przewidywane"



Rys. Okno „predykcja"

W kolumnie „przewidywane wartości" znajdują się prognozy lasu losowego-kopiujemy kolumnę prognoz do arkusza kalkulacyjnego – zakładka „las_losowy – wklejamy wyniki do kolumny E lub F w zależności od badanego horyzontu prognozy (t+1 lub t+7)
W kalkulatorze błędów pojawią się obliczone błędy MAPE% oraz RMSE – kopiujemy wyniki do tabeli z wariantami oraz opisujemy w tabeli parametry danego wariantu. Kolumnę prognoz wklejamy do archiwum wyników.

kolejne kroki: szukamy właściwych parametrów lasu losowego dla których wartości miar błędów w zakresie danych testowych będą najmniejsze.

---

6. Prognozy z wykorzystaniem techniki uczenia maszynowego – wzmacniane gradientowo drzewa decyzyjne dla horyzontu t+1 oraz horyzontu t+7

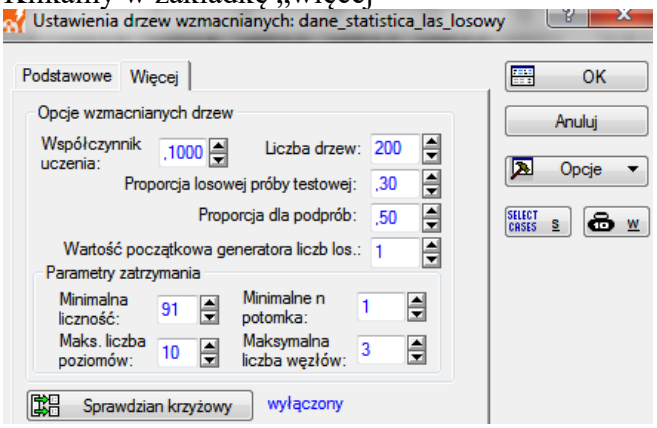
W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 66, przypadków 3652
W arkuszu excel zaznaczamy zakres B1:BH3653 czyli z nagłówkami wybrane dane z zakładki "dane_uczenie_maszynowe" (58 zmiennych (w tym 56 potencjalne zmienne wejściowe), 3652 przypadków)

W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych

menu -> Data Mining -> pod napisem „podstawowe” (górne menu) wybieramy czerwoną ikonę z napisem „wzmacniane drzewa klasyfikacyjne i regresyjne” (środkowa ikona w 1 rzędzie czerwonych ikon)

w oknie wybieramy pozycję -> zadanie regresyjne, klikamy ok.

w zakładce „podstawowe” wybieramy przycisk „zmienne” wskazujemy wyjście modelu (zmienna zależna) oraz zestaw potencjalnych wejść do wyboru przez algorytm w analizie (predyktory ilościowe)
trzymając klawisz CTR można wybrać predyktory ilościowe lub wpisać liczbowo w dolnym oknie.
Klikamy w zakładkę „więcej”



Rys. Okno parametrów drzewa wzmacnianego gradientowo

W zakładce „więcej” wybieramy
a) współczynnik uczenia
b) liczbę drzew
c) parametry zatrzymania (elementy regularyzacji)
Następnie klikamy w przycisk „sprawdzian krzyżowy” w celu wskazania próby uczącej (sygnatura 1) (pozostałe przypadki będą próbą testową) – klikamy w „stan włączony”,

klikamy w „zmienna identyfikująca próby” – wybieramy pozycję pierwszą z listy – kod_danych_losowy_TRE_TEST, klikamy OK.,
następnie podajemy przy polu „kod próby do analizy” liczbę 1 (oznacza to, że próba ucząca to dane z tą sygnaturą), klikamy OK.

Uwagi:
Współczynnik uczenia. Wzmacniane drzewa oblicza ważone "addytywne" rozwinięcia prostych drzew regresyjnych. Waga z jaką do równania predykcji dodawane są kolejne drzewa jest zazwyczaj stała i nazywana jest współczynnikiem uczenia (learning rate) lub parametrem redukcji (shrinkage parameter). Badania empiryczne pokazują, że lepszy model uzyskuje się zazwyczaj, gdy wartość tego parametru wynosi 0,1 lub mniej (wtedy jest lepszą trafność prognostyczną).
Liczba drzew. Tutaj podajemy liczbę składników, które mają być obliczane, tzn. liczbę prostych drzew regresyjnych tworzonych w kolejnych krokach wzmacniania.

Parametry zatrzymania – cztery parametry .
Parametry w tej grupie pozwalają sterować złożonością drzew budowanych w kolejnych krokach wzmacniania. Zalecane jest używanie w miarę prostych drzew w każdym kroku (można np. nie zmieniać ustawienia domyślnego, czyli prostego pojedynczego podziału równoważnego drzewu o 3 węzłach).

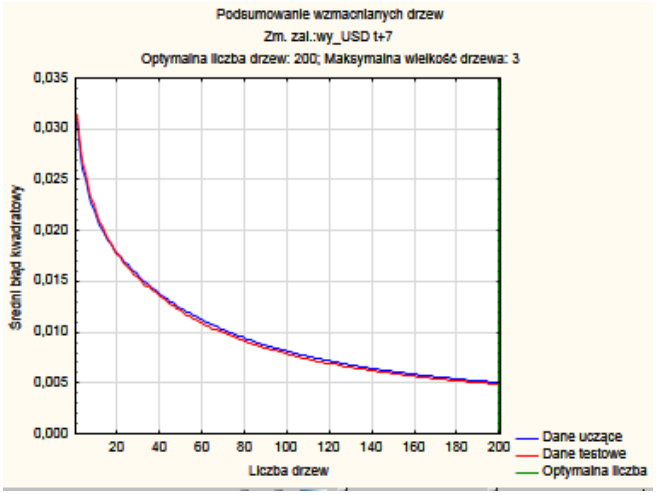
1. Minimalna liczność. Jednym ze sposobów sterowania podziałami jest dopuszczenie do podziałów aż do momentu, gdy wszystkie węzły końcowe będą zawierać nie więcej niż podana minimalna liczność przypadków (obiektów). Tą minimalną liczność można określić właśnie w tym polu.

2. **Minimalna liczność potomka**. Ta opcja mówi, jaka jest najmniejsza dopuszczalna liczność potomka by można było wykonać podział. Parametr Minimalna liczność określa czy w węźle można stosować kolejny podział (jeśli węzeł zawiera dużo przypadków, to tak), natomiast parametr minimalna liczność potomka mówi, że podział można zastosować o ile w wyniku podziału nie otrzymamy potomków zawierających zbyt mało przypadków (mniej niż podana liczba). Opcja jest przydatna, jeśli widzimy, że w analizie tworzone są drzewa, które "wypuszczają małe pączki wzdłuż jednego głównego pnia". Podanie wartości tej opcji większej niż 1 (domyślne) zapobiegnie takim sytuacjom.

3. **Maksymalna liczba poziomów**. Wartość w tym polu określa, ile dopuszczamy poziomów drzewa. Przy każdym podziale sprawdzana jest ogólna liczba poziomów ("głębokość" drzewa mierzona od korzenia) i proces dzielenia zatrzymuje się, jeśli liczba poziomów osiągnie podaną liczbę.

4. **Maksymalna liczba węzłów**. Wartość w tym polu oznacza maksymalną liczbę węzłów każdego z drzew. Przy każdym podziale sprawdzana jest liczba węzłów tworzonego drzewa i jeśli osiągnięta zostanie podana tu wartość, proces podziału jest zatrzymywany. Domyślna wartość 3 oznacza, że w każdym kroku wzmacniania tworzone będzie proste drzewo o jednym podziale (składające się z korzenia i dwóch potomków).

Po zakończeniu procesu w zakładce „podstawowe” klikamy w pole „podsumowanie” – na wykresie będzie podana optymalna liczba drzew. Jeśli wynosi ona tyle ile było drzew to warto zwiększyć liczbę drzew.



Rys. Okno wyników poszukiwania optymalnej liczby drzew

Przycisk „ważność” oraz „wykres słupkowy ważności” pokazują ważność zmiennych wejściowych.

Klikamy w zakładkę „predykcja”, wybieramy próba->testowa, klikamy w przycisk „przewidywane”. W kolumnie „przewidywane wartości” znajdują się prognozy -kopiujemy kolumnę prognoz do arkusza kalkulacyjnego – zakładka „wzmacniane_drzewa_grad – wklejamy wyniki do kolumny E lub F w zależności od badanego horyzontu prognozy (t+1 lub t+7)
W kalkulatorze błędów pojawią się obliczone błędy MAPE% oraz RMSE – kopiujemy wyniki do tabeli z wariantami oraz opisujemy w tabeli parametry danego wariantu. Kolumnę prognoz wklejamy do archiwum wyników.

kolejne kroki: szukamy właściwych parametrów drzew wzmacnianych gradientowo dla których wartości miar błędów w zakresie danych testowych będą najmniejsze.

---

7. Prognozy z wykorzystaniem zespołu różnych predyktorów o odmiennym sposobie działania dla horyzontu t+1 oraz horyzontu t+7

Wybieramy kilka najlepszych modeli (predyktorów) z 5 testowanych (można zrobić kilka wariantów zespołów np. 2, 3, 4 lub 5 modeli jako zespół). Warto nie brać pod uwagę wyraźnie gorszych modeli. Analizę wykonujemy w zakładce „zespół predyktorów” w arkuszu kalkulacyjnym.
Do każdego modelu przypisujemy wagi
a. równe wagi (przykładowo dla 4 modeli każda waga to 0,25)
b. wagi wg zależności biliniowej (im lepszy model tym większa waga) wg formuły poniżej (*k*-liczba modeli w zespole)

$$w_i = \frac{\frac{1}{RMSE_i}}{\sum_{i=1}^{k}(\frac{1}{RMSE_i})}$$

Wybrać końcowo najlepszy model dla horyzontu t+1 oraz horyzontu t+7 spośród modeli pojedynczych oraz modeli zbudowanych z zespołów (nie można wykluczyć, że model pojedynczy okaże się najlepszy)

---

8. Analiza biznesowa efektywności prognoz z wykorzystaniem najlepszego znalezionego modelu dla horyzontu t+1 oraz horyzontu t+7

Dla każdego horyzontu wykonać osobno analizę biznesową z wykorzystaniem najlepszej metody dla horyzontu t+1 oraz horyzontu t+7.

Opis metody dla horyzontu t+1 prognoz.
Przedsiębiorca może kupić towar o wartości 10000 USD w USA płacąc w dolarach. Robi to 549 razy (tyle mamy prognoz dla zakresu testowego). W danym dniu np. 8.10.2010 (pierwszy rekord danych w zakładce „analiza biznesowa") czyli dniu t może zamówić towar lub w dniu t+1 (analiza dla horyzontu t+1 prognoz).
Decyzję podejmuje za każdym razem na dwa sposoby (metody):
a. 549 razy losowo wybiera na moment zakupu dzień t lub dzień t+1
b. 549 razy sugeruje się prognozą czyli zamawia towar wtedy gdy kurs jest niższy w wyniku porównania kursu z dnia t oraz prognozy na dzień t+1 czyli

- kupuje w dniu t (wartość znana) jeśli prognoza na dzień t+1 wskazuje, że kurs będzie większy lub równy w dniu t+1 niż w dniu t
- kupuje w dniu t+1 jeśli prognoza wskazuje, że kurs będzie niższy w dniu t+1 niż w dniu t

Należy dla obu horyzontów policzyć całkowity koszt dla decyzji metodą a) oraz metodą b) oraz obliczyć zysk (o ile wystąpi) uzyskany dzięki metodzie b) wyrażony w złotówkach oraz procentowy.

Dla horyzontu t+7 metoda jest identyczna tylko horyzont prognozy to t+7 czyli zamawia towar albo w dniu t albo w dniu t+7 (podejmuje decyzję w dniu t)

Porównać skuteczność strategii a oraz strategii b jako procent dobrych decyzji z 549 decyzji.

Sprawozdanie: 1 wstęp, 2. Tabelaryczna i graficzna prezentacja wyników, 3. Wnioski