Warsztaty badawcze - Projekt 1

Agata Czajkowska

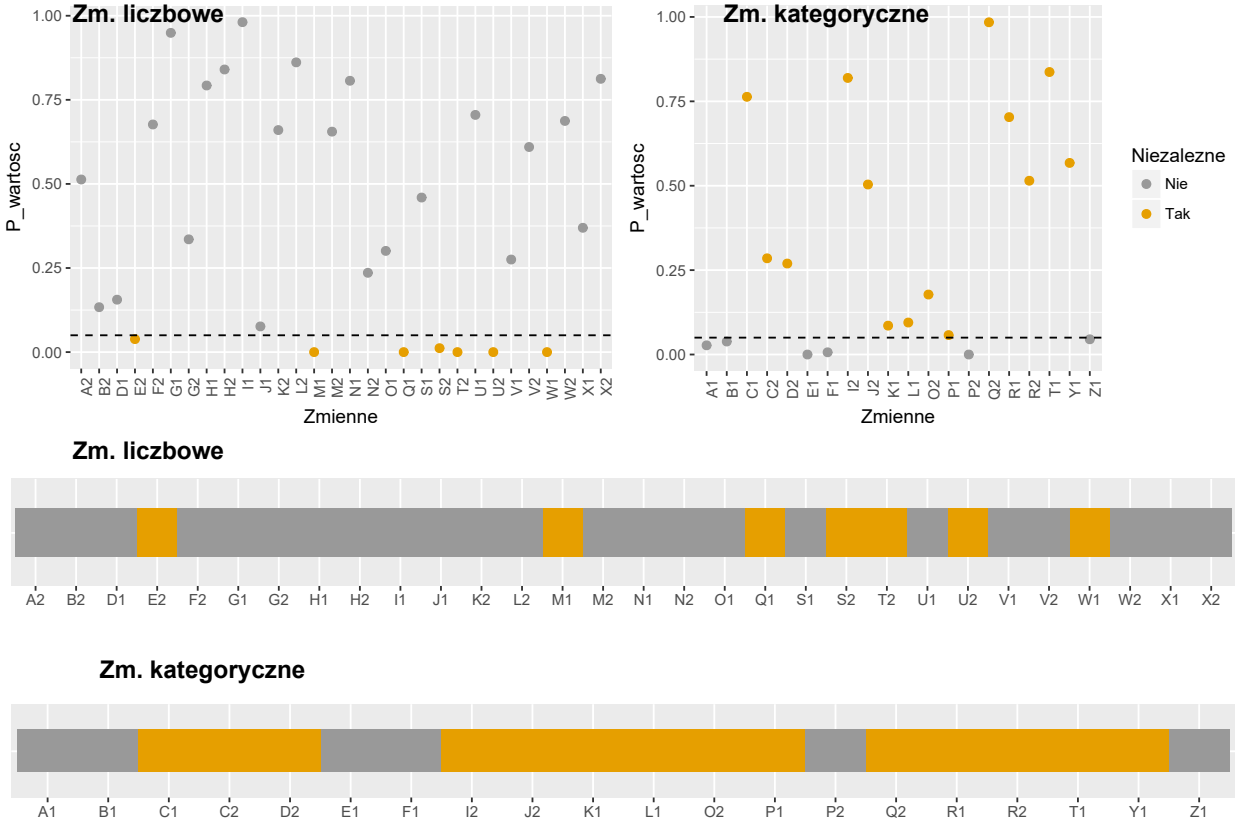
Październik 14, 2017

W projekcie należało dokonać zadania predykcji na sztucznie wygenerowanych danych. Dane składały się z 50 cech i jednej zmiennej zależnej, która mogła przyjąć 2 możliwe wartości "klasa +" lub "klasa -". By w efekcie możliwe było sprawdzenie jakości predykcji dostarczony zbiór treningowy podzielono w sposób losowy na zbiór treningowy i testowy z zachowaniem proporcji 4:1.Wszystkie testy, decyzje o istotności zmiennych wykonano na poziomie istotności $\alpha = 0.05$.

1. Zależność między zmiennymi X a zmienną y

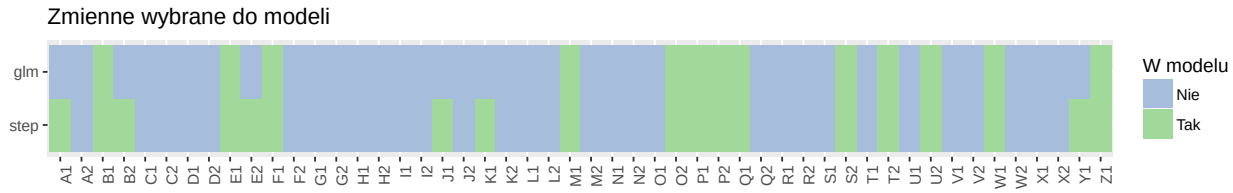
W pierwszym kroku wykonano działania mające na celu wykrycie zależności pomiędzy zmiennymi objaśniającymi X a zmienną objaśnianą y. Analizy dokonano odzielnie dla zmiennych kategorycznych i zmiennych liczbowych.

- Dla zmiennych kategorycznych wykonywano chisq.test(). Hipoteza zerowa zmienna X_i i zmienna y są niezależne.
- Zależność między zmiennymi liczbiwymi a zmienną y weryfikowano przy pomocy anovy. Hipoteza zerowa: zmienna X_i i zmienna y są zależne.



2. Selekcja zmiennych

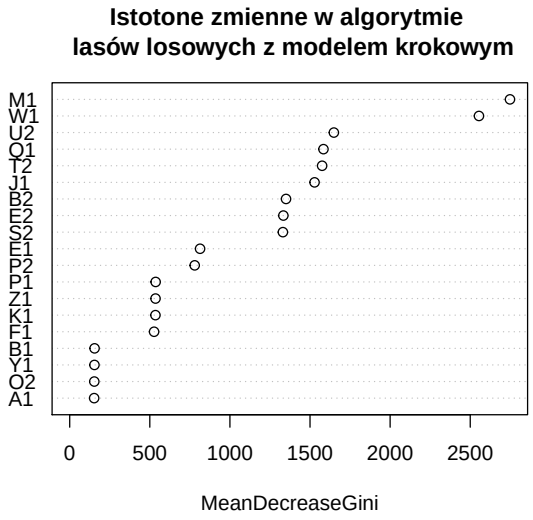
Selekcji zmiennych dokonano na dwa sposoby : za pomocą regresji logistycznej i przy użyciu metody krokowej z kierunkiem wstecz i karą AIC.W efekcie utworzono 2 modele. Zostały wybrane do nich następujące zmienne.

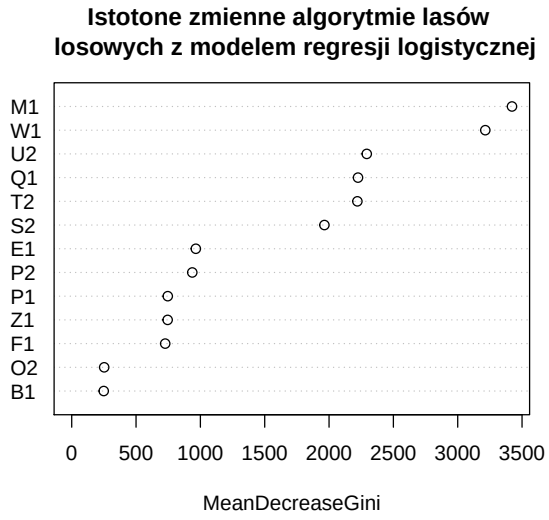


Procedura step wybiera większy model niż regresja logistyczna. Wszystkie zmienne wybrane w modelu regresji liniowej zostały też wybrane przez procedurę krokową. Ponad to istnieją zmienne wybrane przez modele, dla których nie wykryto zależności między daną zmienną a zmienną objaśnianą np.M1. Istnieją też takie zmienne, dla których wykryto zależności w poprzednim kroku analizy, a żaden ze sposobów selekcji zmiennych nie wybrał danej zminnej do modelu np.V1.

3. Klasyfikacja

Klasyfikacji dokonywano za pomocą dwóch metod : regresji logistycznej i z użyciem lasów losowych. Dla lasów losowych ilość drzew wynosiła 1000. Analizy dokonano dla każdej z metod na obu modelach wyznaczonych w poprzednim kroku.





W obu przypadkach zmienne $M1, W1, U2, Q1$ uznano za najbardziej istotne w podanej kolejności. Następnie w większym modelu za bardziej istotną uznano zmienne $J1, B2, E2$ niż zmienną $S1$.

Jako miarę jakości klasyfikacji wybrano precyzję dla 10% predykcji z najwyższymi prawdopobieństwami przynależności do klasy "+" . Otrzymano następujące wyniki:

Metoda	Model	Wynik
Lasy losowe	Krokowy	0.882
Lasy losowe	Regresja logistyczna	0.868
Regresja logistyczna	Krokowy	0.746
Regresja logistyczna	Regresja logistyczna	0.751

Najlepszą skuteczność rzędu 88% dał model z otrzymany z procedury step a którym użyto algorytmu Rforest. Najmniejszą rzędu 75 % dał model z otrzymany z procedury step na którym użyto regresji logistycznej.