Opis dotychczasowej analizy o restauracjach

Zgodnie z Pani sugestią z naszego ostatniego spotkania skupiłem się na ocenie co bardziej wpływa na rozmieszczenie restauracji- biznesy w pobliżu czy liczba mieszkańców. Podszedłem do tego od strony modelowania, i za cel przyjąłem przewidzenie czy w danym obszarze jest jakakolwiek restauracja. Do rastra dokleiłem zmienne oznaczające liczbę przystanków i długość dróg w obszarze. Dokładniejszy opis użytych technik znajduje się poniżej.

Modele które wyestymowałem mają bardzo dużą dokładność, nawet przy użyciu tylko jednego predyktora. Żadna ze zmiennych nie wyróżnia się zbytnio- różnice w wynikach przy użyciu różnych podzbiorów zmiennych są rzędu 1%. W kontekście pytania badawczego wniosek jest taki, że różnica pomiędzy wpływem ilości biznesów i ilości mieszkańców jest niewielka. Co więcej, użycie tylko zmiennych dotyczących infrastruktury (ilość dróg i przystanków) daje porównywalnie dobre predykcje.

Pytania

1. Użyłem danych o biznesach które Pani przesłała, ale w API z którego korzystałem wyczerpałem limit na geokodowanie w tym miesiącu. Dlatego wybrałem tylko subset 15 tys. biznesów. Myślę że to całkiem sporo i dodawanie reszty nie jest konieczne, jeżeli jednak posiada Pani zakodowane te adresy to byłbym wdzięczny za udostępnienie.
2. Czy używanie I Morana jest usprawiedliwione dla zmiennych kategorycznych (konkretnie czy w danym obszarze jest restauracja)? Jedyne co o tym znalazłem to wpis na forum, w którym ktoś przekonywał że nie ma przeciwwskazań ale bez podania argumentów.
3. Używając I Morana otrzymałem wartości dla zmiennej celu 0.46. Dalej do estymacji modeli utworzyłem zmienne odpowiadające *spatial lag* zmiennej celu i zmiennych objaśniających (np. neighbours_restaurant_count= suma ilości restauracji w sąsiednich obszarach). Czy, jako że nie ma bardzo silnej korelacji przestrzennej, powinienem usunąć te zmienne z modeli?
4. Czy mogłaby Pani polecić jakieś artykuły o estymacji modeli przestrzennych dla klasyfikacji? Wszystko co znajdowałem dotyczyło regresjii.

Key findings from Exploratory Data Analysis:

- A restaurant in the area positively correlates with population density, business number, total roads length in the area and bus stops number.
- Both restaurants and business count in the area are highly correlated (Pearson's correlation= +0.71), while population density is weakly positively correlated with restaurants number (Pearson's correlation= +0.34)
- Restaurants number and business number are highly positively skewed, 67% of areas don't have restaurants and 40% don't have businesses (Using ~10% of businesses dataset, so this value is probably smaller in reality).

- Spatial autocorrelation measured by Moran's I is 0.46 for restaurant count, 0.58 for population density and 0.15 for business count.

Description of analysis

Main goal of this study was to check what influences the location of restaurants in given area more: population density or business number.

Because of the fact that in most grid cells the number of restaurants was 0, I have decided to perform classification rather than regression (where the target variable is if there is a restaurant in given area).

General approach was to estimate some models using all avaliable variables, and then check their predictive value. Specifically, approaches I have tested are as follows:

- **Method 1:** Estimate Random Forest model on all variables and then check variable importance (as specified in randomForest package description here).
- **Method 2:** Estimate Random Forest and Logistic Regression models 3 times each, first time using all variables except population density, second- except number of businesses, and third- except both population density and number of businesses. Third model serves as a baseline. Then I made prediction on held-out data using every model and checked which variable (population or businesses) gives bigger improvement compared to the baseline. I used AUC as comparison metric.
- **Method 3:** Estimate 6 models as above, but use resampling of observations to get more accurate AUC estimates as described here and here

The results of analysis are as follows:

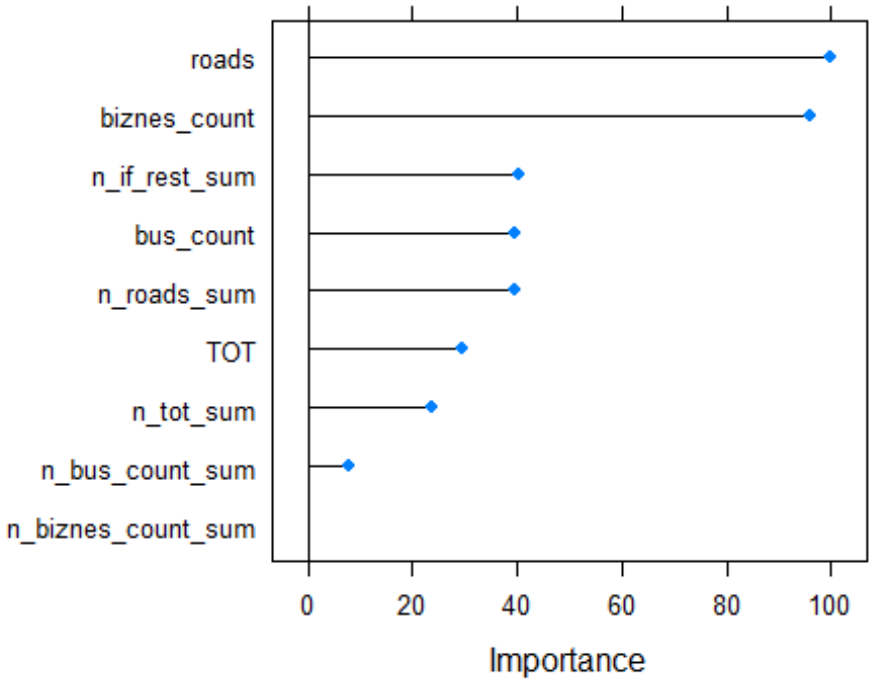
Method 1

The most important predictor is a variable indicating the sum of areas around, in which restaurant is present. Another important variable is total length of roads in the area. Another variables have smaller importance regarding prediction. Specifically, the number of businesses in area has 90% of predictive power of the strongest variable. The variable indicating population density in the area has only 15% of best variable predictive power.

Using this method, one can conclude that number of businesses in the area has bigger influence on the decision to open a restaurant than population density in given area.

(labels on the plot: TOT- total population density, variables starting with n- neighbouring regions)

```
plot(caret::varImp(model_rf_all))
```



Method 2

It should be stated that both models perform pretty much the same, and AUC value for both Random Forest and Logistic Regression is in the range (0.88, 0.91). This could mean that business count and population density aren't important predictors.

To assess the results, best way is to compare results for Random Forest and Logistic Regression separately. For RF, best score is obtained by using all variables (AUC 0.91). Second best model is the one not containing population density (0.905). Next are model without businesses and population (0.894), and the last one- not containing businesses (0.889). This is a yet another indicator that business presence in the area is more important to restaurants location than population density. Although the differences in obtained results are very small and could occur by coincidence.

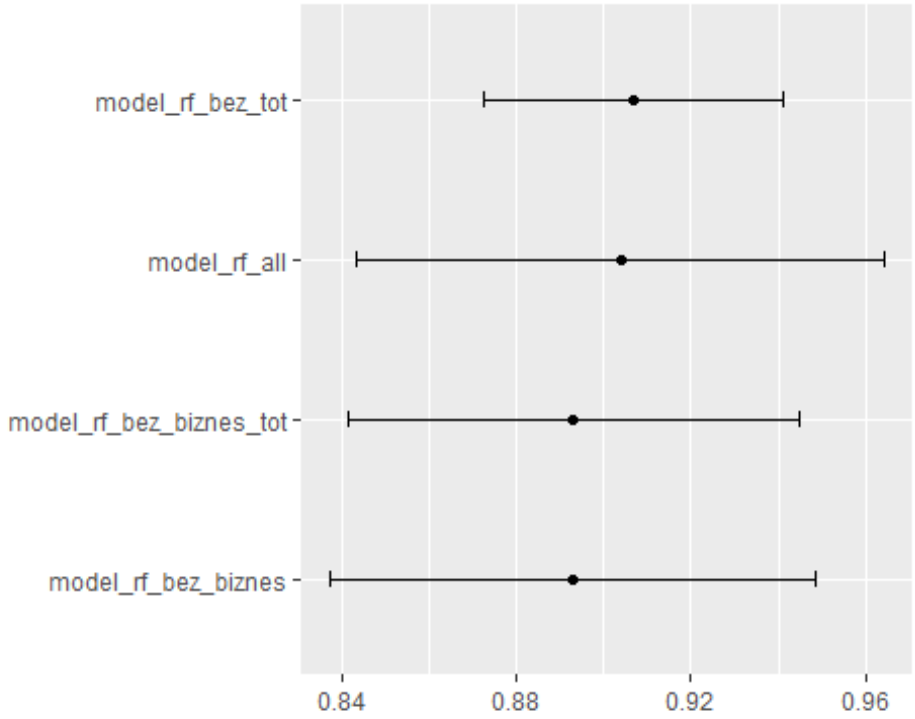
For Logistic Regression, the results are different. The best model is the one not containing business count in the area (AUC 0.902). Next is the one without businesses and population (0.901), including all variables (0.900) and the worst one- without population density (0.897). Same as in Random Forest, difference between models are small.

Method 3

AUC estimations using resampling method for all tested are shown on a plot below. It can be concluded that there is no significant difference in AUC estimates between models, and thus both business count and population density are equally important predictors.

Resamples results plot for random forest model:

```
ggplot2::ggplot(resamps_rf)
```



Details

To take spatial dimension into account, I have inluded variables containing information about sum of independent variable in the neighbouring observations (Areas are neighbours when they have at leas 1 common vertex or edge).

To fit the Random Forest model 1/5 of all data was held out as test set. On the training set 4-fold cross-validation was performed. The folds were selected by binding adjacent Warsaw districts together, not randomly. This was to ensure that no leaks of the training data to test data was created. This process was similar to one described here.

Variables used

One observation means one 1km x 1km cell. Variables include:

- Restaurants count in the area- points data binned to cells. In prediction I have used varaible if_restaurant which is "y" when there is a restaurant in the area
- Businesses count in the area- points data binned to cells.
- Population count in the area - originally in 1km x 1km grid.
- Total length of roads in the cell- taken from OpenStreetMap
- Bus stops count- points data binned to cells, taken from OpenStreetMap

To above variables I have added their spatially lagged versions:

- Count of adjacent cells in which there is restaurant

- Sum of businesses count in adjacent cells
- Sum of population count in adjacent cells
- Sum of roads length in adjacent cells
- Sum of bus stops count in adjacent cells