Predicting Views, Reactions and Conversions

A Preliminary Analysis via Machine Learning

Patrik Jankovič, Štefan Lyócsa & Miroslav Štefánik

Institute of Economic Research
Slovak Academy of Sciences
Šancova 56, Bratislava, Slovakia

September 24, 2020

Goal

Identify **models** and **drivers** of attractiveness of **job offers** at 'profesia' - leading job advertising portal in Slovakia.

Motivation

- Can we help advertisers to make job offerings more interesting?
- What makes a job offering more attractive?
- Are drivers business area specific or geographically specific?
- What is the expected range of interest generated by a given job offer?
- Can we increase (hence monetize) views, reactions, conversions?

Challenges

We use data from 2019.

1. Fairly large (final) data set: 249812 (obs.) $\times$ 671 (variables).
2. Heterogeneous data (numbers & text, diacritics).
3. No template $\rightarrow$ many (feasible) alternatives to analysis.
4. Difficult to interpret models & variables.
5. Computational & memory intensity.

50% of time devoted to data management. Some data categories:

- 115 regions (diacritics..).
- 52 business areas.
- 25 job benefits.
- 37 calendar effects (weekdays, months, holidays).
- Position names - number and length of words.
- Salary information: text, numbers, formats,...makes analysis difficult **to code** and **to reproduce**).

ReactDum:

- 0 if job offer had 0 reactions.
- 1 if job offer had > 0 reactions.

81.5% job offers received a reaction(s).

A recent study of Bastani et al., (2019) proposes a two-stage analysis. In our context, **first**, we should train one model on job offers with no reaction at all (to identify 'unpopular job offers') and **second**, we should train one model on job offers with positive reactions.

Reactions a positive integer of reactions for a job offer. With median at 7.0, mean at 16.2 and SD 33.7, reactions appear to be **extremely volatile**.

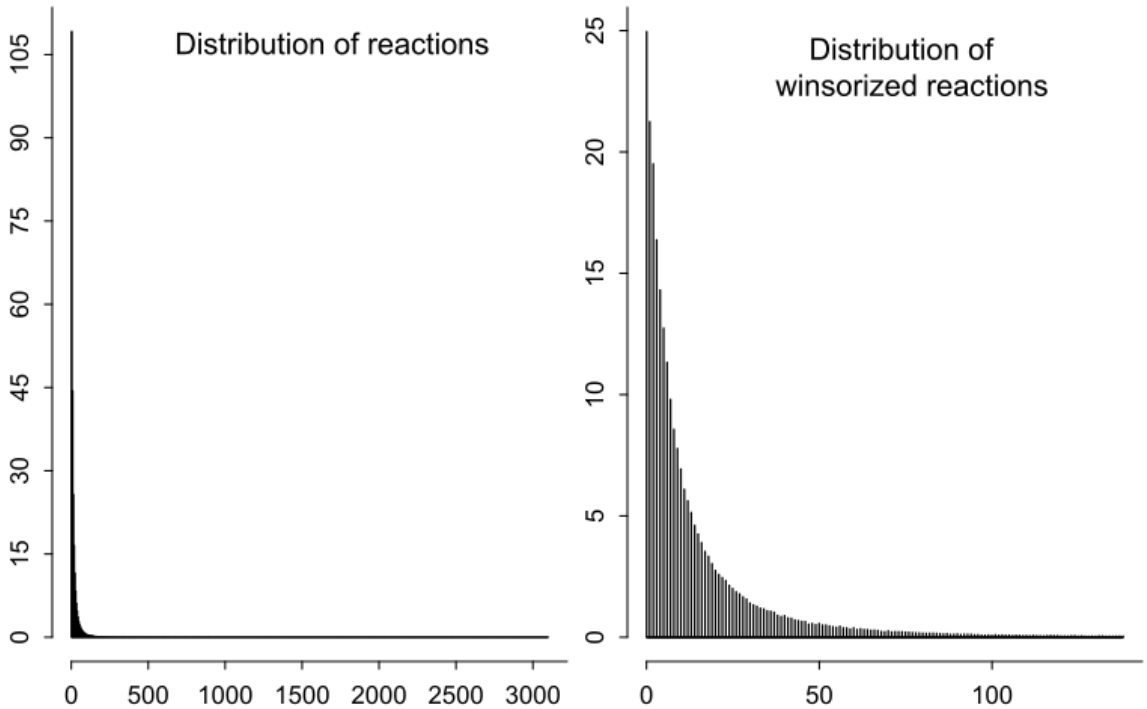

Figure 1: Reactions

Distribution suggests (power-law?) tendency to **extremes**.

Views a positive integer of job offer views. With median at 450, mean at 663 and SD 845.5, views are also **substantially volatile**. Views and reactions are correlated (0.76 Spearman's).

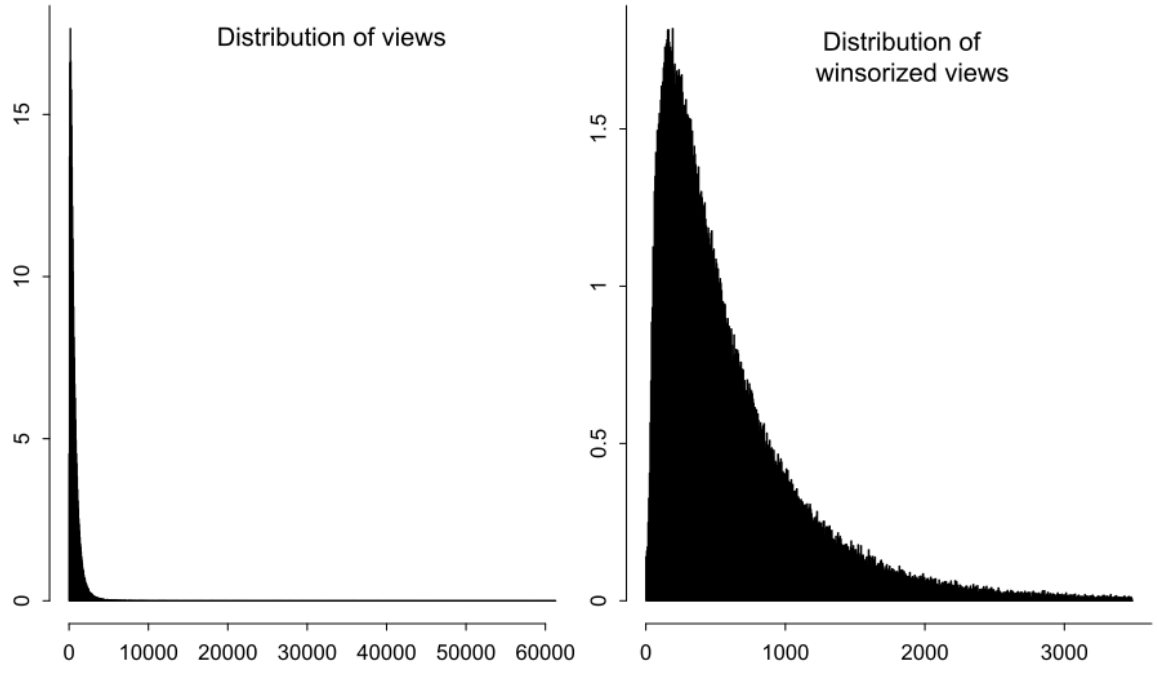

Figure 2: Views

Conversions are defined as:

$$\frac{Reactions}{Views} \tag{1}$$

Summary statistics: median 0.0158, mean 0.0215 and SD 0.026.

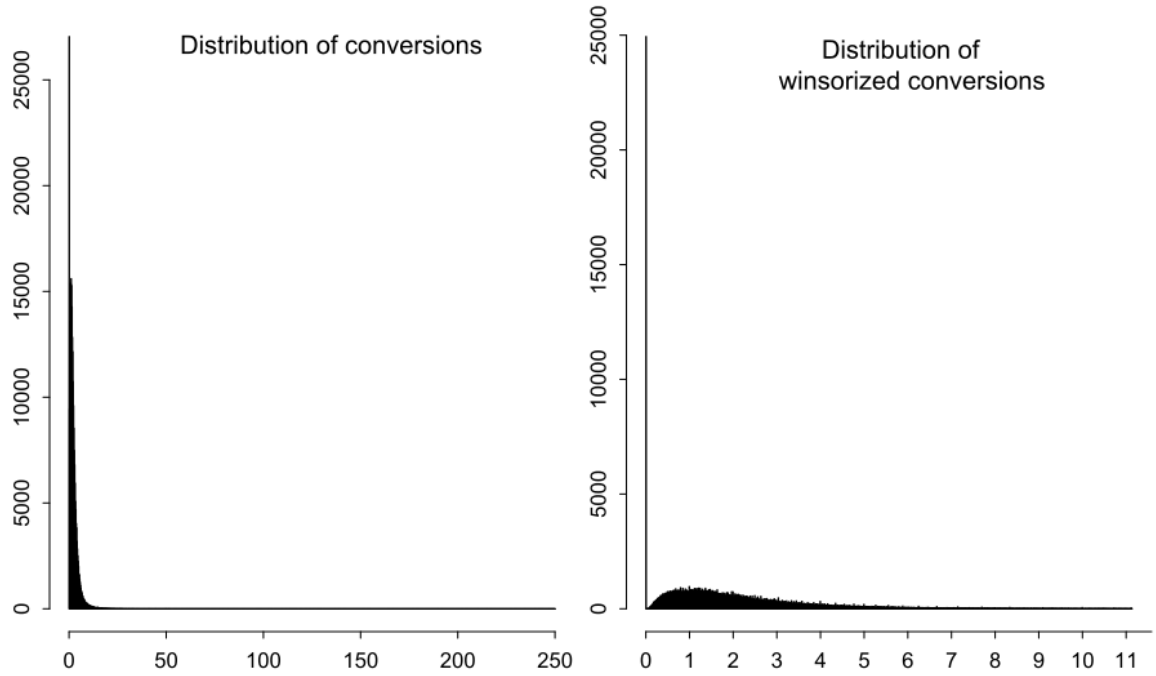

Figure 3: Conversions

Key concepts at glance

- Analysis run on given region and business area combination, e.g. 'Telecommunications in Bratislava'.
- Variable importance and model parameter tuning on the testing data set.
- Model verification on the validation data set.

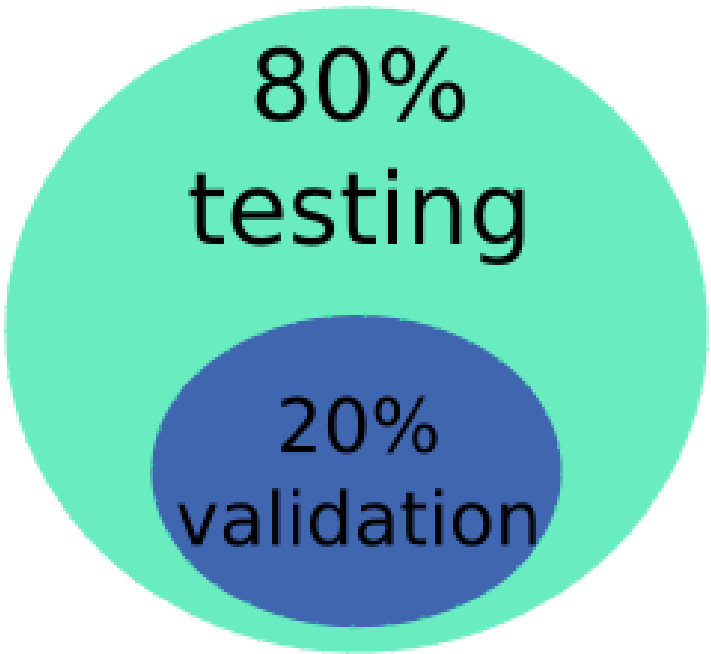

Figure 4: Data sample stratification

Models

Two-stage analysis of Bastani et al., (2019). Consider modeling **conversions** for job offers **in Bratislava** in the Business Area of **Information Technologies**.

1. We model *ReactDum* ($\approx 31.5\%$ no reaction/conversion) using the testing data set: Logistic regression, LASSO, Ridge and Random Forest.
2. We model *Conversions* using the training data set and only **with** *Reactions* > 0: OLS, LASSO, Ridge and Random Forest.
3. Given model from step 1, we predict job offers that are expected to have 0 or *more* Reactions.
4. Given model from step 2, we predict conversions only for job offers with expected *Conversion* > 1.

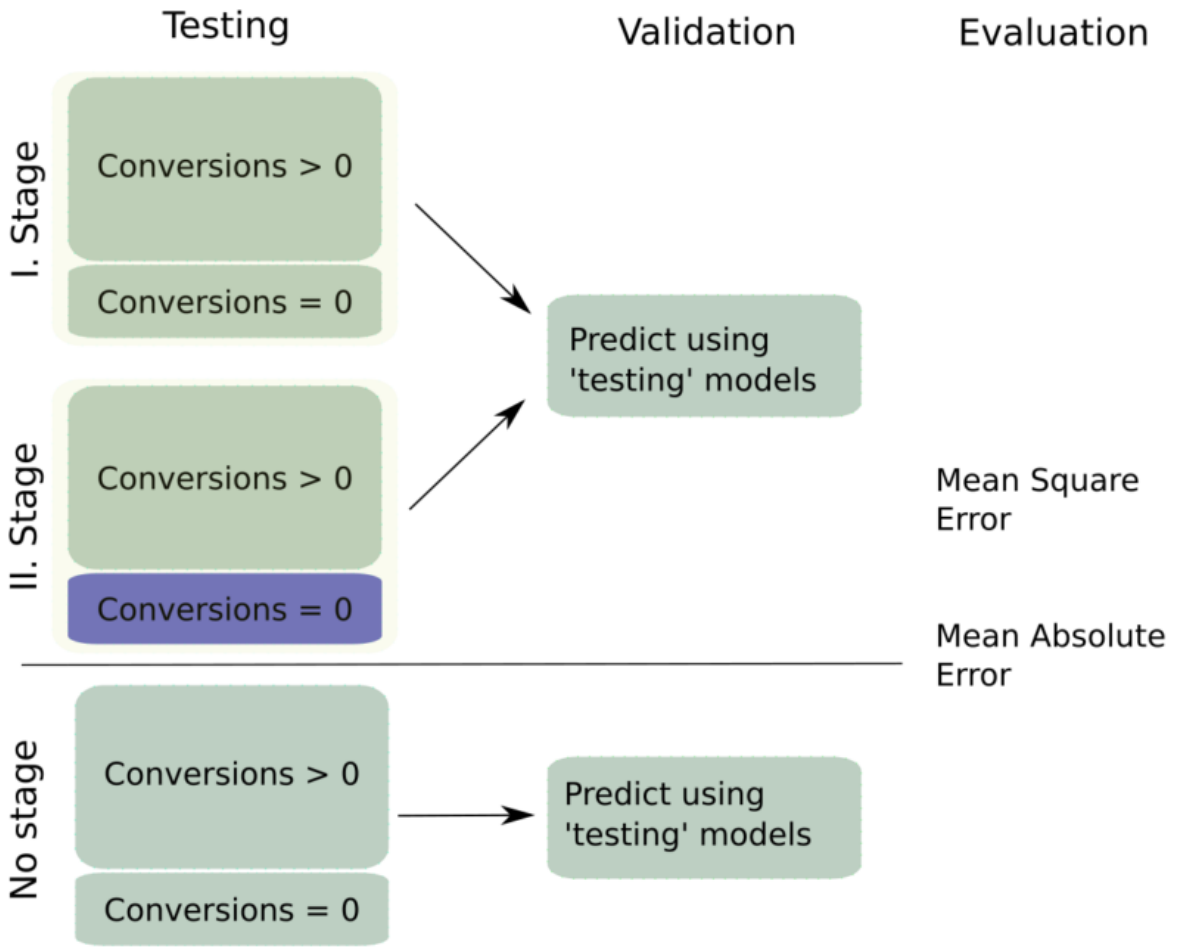

Figure 5: Approach to predictions

Statistical comparison of forecasts via the **model confidence set** of Hansen et al., (2011).

Conversion averaged at 0.029, median 0.019 and SD 0.0048. Prediction errors are:

Model	MSE		MAE		Median AE	3rd Quart.
Panel A: Unconditional benchmark models as sample averages across job offers						
All job offers	0.0049	0.0%	0.021	0.0%	0.013	0.022
Job offers in region	0.0049	-0.4%	0.025	22.4%	0.021	0.030
Job offers in business area	0.0049	-0.6%	0.021	1.4%	0.014	0.024
Job offers in region & business area	0.0048	-1.1%	0.022	5.6%	0.015	0.024
Panel B: Two-stage models						
SMOTE-LR-OLS	0.0049	-0.5%	0.023	11.3%	0.015	0.028
WGT-LR-OLS	0.0049	0.2%	0.023	12.9%	0.016	0.028
SMOTE-LASSO-MIN	0.0049	-0.2%	0.023	11.3%	0.016	0.027
WGT-LASSO-MIN	0.0049	-0.5%	0.023	12.5%	0.016	0.028
SMOTE-RIDGE-MIN	0.0048	-1.4%	0.023	11.2%	0.016	0.027
WGT-RIDGE-MIN	0.0049	-0.4%	0.023	12.4%	0.016	0.028
SMOTE-RF	**0.0046**	-5.5%	0.020	-4.4%	0.013	0.023
WGT-RF	0.0047	-4.6%	0.020	-2.9%	0.013	0.024
Panel C: One-stage models						
OLS	**0.0047**	-4.7%	0.020	-1.5%	0.013	0.023
LASSO-MIN	**0.0047**	-4.7%	0.020	-1.6%	0.013	0.023
RIDGE-MIN	**0.0047**	-4.6%	0.020	-1.5%	0.013	0.023
RF	**0.0046**	**-6.3%**	**0.018**	**-10.6%**	0.011	0.020

Table 1: Banking and Money in Bratislava: Conversions

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Conversion for 50% lowest [for 0 if dummy]	Conversion for 50% highest [for 1 if dummy]	t test [p-value]
Salary information (clear)	[1 - yes, 0 - no]	0.0002098	0.043	0.025	0.000
Number of letters in the title	log	0.0001793	0.029	0.025	0.000
University - bachelor	[1 - yes, 0 - no]	0.0001577	0.029	0.022	0.000
Middle school graduate	[1 - yes, 0 - no]	0.0001267	0.023	0.031	0.000
Number of words in the title	log	0.0001183	0.027	0.027	0.888
University - master	[1 - yes, 0 - no]	0.0000643	0.028	0.020	0.000
2 to 3 words in the title	[1 - yes, 0 - no]	0.0000417	0.026	0.030	0.001
March	[1 - yes, 0 - no]	0.0000413	0.026	0.039	0.000
Suitable for graduates	[1 - yes, 0 - no]	0.0000282	0.027	0.029	0.009
April	[1 - yes, 0 - no]	0.0000265	0.025	0.044	0.000

Table 2: Banking and Money in Bratislava: Conversions

Reactions averaged at 13.69, median 6.00 and SD 23.02. Prediction errors are:

Model	MSE		MAE		Median AE	3rd Quart.
Panel A: Unconditional benchmark models as sample averages across job offers						
Mean.All	536.0	0.0%	14.5	0.0%	12.2	15.2
Mean.Region	536.2	0.0%	14.5	0.1%	12.2	15.2
Mean.Business.Area	532.0	-0.7%	14.0	-3.4%	11.2	14.2
Mean.Region.Business.Are	529.7	-1.2%	13.4	-7.5%	11.0	13.0
Panel B: Two-stage models						
SMOTE-LR-OLS	499.8	-6.8%	12.9	-10.9%	9.0	16.7
WGT-LR-OLS	507.3	-5.3%	12.7	-12.2%	8.0	16.5
SMOTE-LASSO-MIN	499.4	-6.8%	12.9	-10.9%	9.2	16.5
WGT-LASSO-MIN	508.0	-5.2%	12.7	-12.4%	8.6	16.2
SMOTE-RIDGE-MIN	501.2	-6.5%	12.9	-10.6%	9.0	16.8
WGT-RIDGE-MIN	507.8	-5.3%	12.7	-12.3%	8.0	16.5
SMOTE-RF	438.5	-18.2%	11.8	-18.3%	7.7	14.3
WGT-RF	433.8	-19.1%	11.4	-21.6%	6.8	14.1
Panel C: One-stage models						
OLS	476.2	-11.2%	12.4	-14.6%	8.5	14.6
LASSO-MIN	477.0	-11.0%	12.4	-14.4%	8.5	14.2
RIDGE-MIN	476.7	-11.1%	12.4	-14.3%	8.6	14.6
RF	**420.7**	**-21.5%**	**11.1**	**-23.6%**	6.4	12.3

Table 3: Banking and Money in Bratislava: Reactions

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Reaction for 50% lowest [for 0 if dummy]	Reaction for 50% highest [for 1 if dummy]	t test [p-value]
Number of letters in the title	log	345.30	17.14	10.99	0.000
4 to 6 words in the title	[1 - yes, 0 - no]	250.53	11.85	19.78	0.000
University - master	[1 - yes, 0 - no]	236.14	15.21	11.10	0.000
Middle school graduate	[1 - yes, 0 - no]	197.33	13.09	15.06	0.001
Number of words in the title	log	174.15	13.98	13.99	0.992
Salary information (clear)	[1 - yes, 0 - no]	105.18	23.16	12.85	0.000
University - bachelor	[1 - yes, 0 - no]	86.14	14.90	8.68	0.000
University - student	[1 - yes, 0 - no]	40.22	13.63	22.56	0.000
Suitable for graduates	[1 - yes, 0 - no]	26.78	13.90	14.39	0.469
2 to 3 words in the title	[1 - yes, 0 - no]	25.29	14.64	12.57	0.001

Table 4: Banking and Money in Bratislava: Reactions

Views averaged at 448, median 301 and SD 441. Prediction errors are:

Model	MSE		MAE		Median AE	3rd Quart.
Panel A: Unconditional benchmark models as sample averages across job offers						
All job offers	241207	0.0%	410.7	0.0%	415.0	523.0
Job offers in region	195732	-18.9%	296.6	-27.8%	231.6	325.6
Job offers in business area	216043	-10.4%	370.8	-9.7%	355.1	460.1
Job offers in region & business area	195030	-19.1%	302.8	-26.3%	245.9	339.1
Panel B: Two-stage models						
SMOTE-LR-OLS	199695	-17.2%	318.7	-22.4%	235.9	384.0
WGT-LR-OLS	209610	-13.1%	328.9	-19.9%	242.0	394.0
SMOTE-LASSO-MIN	201360	-16.5%	320.6	-21.9%	239.4	384.0
WGT-LASSO-MIN	210363	-12.8%	330.1	-19.6%	246.7	394.9
SMOTE-RIDGE-MIN	197925	-17.9%	318.3	-22.5%	236.8	379.0
WGT-RIDGE-MIN	210726	-12.6%	330.6	-19.5%	247.7	395.6
SMOTE-RF	**130818**	-45.8%	248.7	-39.4%	184.0	303.8
WGT-RF	137844	-42.9%	256.3	-37.6%	185.3	308.9
Panel C: One-stage models						
OLS	**149672**	-37.9%	260.7	-36.5%	190.2	309.3
LASSO-MIN	**150052**	-37.8%	261.3	-36.4%	193.2	307.8
RIDGE-MIN	**150004**	-37.8%	261.5	-36.3%	193.1	307.5
RF	**114662**	**-52.5%**	**219.1**	**-46.7%**	146.6	256.2

Table 5: Banking and Money in Bratislava: Views

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Views for 50% lowest [for 0 if dummy]	Views for 50% highest [for 1 if dummy]	t test [p-value]
Number of letters in the title	log	70691	510.7	374.5	0.000
Number of words in the title	log	51259	417.3	463.0	0.000
Elementary school	[1 - yes, 0 - no]	47322	419.6	1781.3	0.000
2 to 3 words in the title	[1 - yes, 0 - no]	32688	384.0	595.5	0.000
Personal agency	[1 - yes, 0 - no]	23843	424.5	1738.5	0.000
Middle school graduate	[1 - yes, 0 - no]	14893	467.0	409.7	0.000
University - master	[1 - yes, 0 - no]	11258	460.1	395.8	0.000
Suitable for graduates	[1 - yes, 0 - no]	11246	430.7	485.1	0.001
English language	[1 - yes, 0 - no]	8384	442.0	434.8	0.523
Slovak language	[1 - yes, 0 - no]	7915	434.0	442.2	0.463

Table 6: Banking and Money in Bratislava: Views

Conversion averaged at 0.020, median 0.017 and SD 0.020.

Model	MSE		MAE		Median AE	3rd Quart.
Panel A: Unconditional benchmark models as sample averages across job offers						
All job offers	0.00039	0.0%	0.0131	0.0%	0.011	0.017
Job offers in region	0.00039	-0.3%	0.0129	-1.1%	0.011	0.017
Job offers in business area	0.00040	3.8%	0.0140	7.4%	0.013	0.019
Job offers in region & business area	0.00039	-0.1%	0.0130	-0.2%	0.011	0.017
Panel B: Two-stage models						
SMOTE-LR-OLS	0.00053	36.3%	0.0167	27.6%	0.013	0.023
WGT-LR-OLS	0.00053	37.3%	0.0167	27.9%	0.013	0.022
SMOTE-LASSO-MIN	0.00052	33.2%	0.0162	24.2%	0.013	0.022
WGT-LASSO-MIN	0.00053	35.2%	0.0164	25.3%	0.013	0.022
SMOTE-RIDGE-MIN	0.00052	33.8%	0.0164	25.4%	0.013	0.022
WGT-RIDGE-MIN	0.00053	36.6%	0.0166	26.9%	0.014	0.022
SMOTE-RF	0.00038	-3.3%	0.0137	5.2%	0.011	0.019
WGT-RF	0.00037	-5.4%	0.0135	3.7%	0.011	0.019
Panel C: One-stage models						
OLS	0.00037	-4.1%	0.0133	1.6%	0.010	0.018
LASSO-MIN	**0.00036**	**-7.3%**	**0.0128**	**-2.2%**	0.010	0.018
RIDGE-MIN	**0.00036**	**-6.8%**	**0.0129**	**-1.4%**	0.010	0.019
RF	**0.00032**	**-18.4%**	**0.0125**	**-4.7%**	0.009	0.017

Table 7: Hospitals, ambulance, doctors in Bratislava: Conversions

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Conversion for 50% lowest [for 0 if dummy]	Conversion for 50% highest [for 1 if dummy]	t test [p-value]
Number of letters in the title	log	0.0000780	0.022	0.021	0.609
Midle school - graduate	[1 - yes, 0 - no]	0.0000660	0.018	0.028	0.000
Number of words in the title	log	0.0000655	0.021	0.022	0.532
Midle school - no graduate	[1 - yes, 0 - no]	0.0000640	0.025	0.016	0.000
7 to 10 words in the title	[1 - yes, 0 - no]	0.0000445	0.021	0.026	0.071
7 to 10 unique words in the title	[1 - yes, 0 - no]	0.0000345	0.021	0.026	0.051
Suitable for graduates	[1 - yes, 0 - no]	0.0000170	0.024	0.018	0.000
Elementary school	[1 - yes, 0 - no]	0.0000167	0.022	0.017	0.000
2 to 3 words in the title	[1 - yes, 0 - no]	0.0000159	0.021	0.024	0.010
University - master	[1 - yes, 0 - no]	0.0000119	0.021	0.030	0.000

Table 8: Hospitals, ambulance, doctors in Bratislava: Conversions

Reactions averaged at 20.66, median 10.00 and SD 33.28. Prediction errors are:

Model	MSE		MAE		MedAe	Q3Ae
Panel A: Unconditional benchmark models as sample averages across job offers						
All job offers	**1125.3**	**0.0%**	**17.2**	**0.0%**	11.2	15.2
Job offers in region	1110.2	-1.3%	18.0	4.6%	12.5	16.5
Job offers in business area	1134.9	0.9%	16.9	-1.7%	10.2	14.2
Job offers in region & business area	1106.6	-1.7%	19.3	12.3%	14.7	19.7
Panel B: Two-stage models						
SMOTE-LR-OLS	1139.3	1.2%	19.1	10.9%	13.0	23.0
WGT-LR-OLS	1145.2	1.8%	19.1	10.9%	12.9	23.0
SMOTE-LASSO-MIN	1121.5	-0.3%	18.7	8.9%	13.5	20.0
WGT-LASSO-MIN	1124.2	-0.1%	18.7	8.6%	13.6	20.1
SMOTE-RIDGE-MIN	1118.4	-0.6%	18.7	8.5%	13.9	20.0
WGT-RIDGE-MIN	1155.1	2.7%	18.9	10.0%	13.5	21.0
SMOTE-RF	985.2	-12.5%	18.2	5.8%	12.8	21.1
WGT-RF	**989.4**	**-12.1%**	**17.6**	**2.3%**	11.8	21.0
Panel C: One-stage models						
OLS	1105.4	-1.8%	19.2	11.5%	13.4	21.8
LASSO-MIN	1085.4	-3.5%	18.9	9.6%	13.6	19.7
RIDGE-MIN	1086.2	-3.5%	19.0	10.3%	13.8	19.3
RF	**973.1**	**-13.5%**	**17.6**	**2.4%**	11.9	20.2

Table 9: Hospitals, ambulance, doctors in Bratislava: Reactions

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Reactions for 50% lowest [for 0 if dummy]	Reactions for 50% highest [for 1 if dummy]	t test [p-value]
Number of letters in the title	log	208.4	22.70	20.74	0.160
Number of words in the title	log	114.6	22.40	21.04	0.327
University - master	[1 - yes, 0 - no]	69.5	23.49	12.21	0.000
Middle school graduate	[1 - yes, 0 - no]	63.6	18.57	24.49	0.000
English language	[1 - yes, 0 - no]	62.4	20.88	27.54	0.000
Slovak language	[1 - yes, 0 - no]	52.8	27.41	20.89	0.000
University - bachelor	[1 - yes, 0 - no]	41.5	20.86	27.31	0.001
2 to 3 words in the title	[1 - yes, 0 - no]	37.9	21.17	22.76	0.287
1 word in the title	[1 - yes, 0 - no]	27.2	21.37	24.15	0.298
Suitable for graduates	[1 - yes, 0 - no]	22.8	22.11	21.29	0.552

Table 10: Hospitals, ambulance, doctors in Bratislava: Reactions

Views averaged at 1003, median 829 and SD 765. Prediction errors are:

Model	MSE		MAE		Median AE	3rd Quart.
Panel A: Unconditional benchmark models as sample averages across job offers						
Mean.All	700333	0.0%	510.9	0.0%	320.5	523.8
Mean.Region	596378	-14.8%	491.7	-3.8%	369.5	567.5
Mean.Business.Area	752767	7.5%	532.8	4.3%	310.5	590.4
Mean.Region.Business.Are	585166	-16.4%	505.9	-1.0%	382.5	625.7
Panel B: Two-stage models						
SMOTE-LR-OLS	722902	3.2%	579.6	13.4%	409.3	735.1
WGT-LR-OLS	758383	8.3%	594.7	16.4%	422.6	744.0
SMOTE-LASSO-MIN	707830	1.1%	566.2	10.8%	404.8	711.1
WGT-LASSO-MIN	703922	0.5%	565.0	10.6%	404.8	712.6
SMOTE-RIDGE-MIN	738717	5.5%	571.7	11.9%	393.9	699.9
WGT-RIDGE-MIN	709313	1.3%	570.4	11.7%	404.2	710.0
SMOTE-RF	509135	-27.3%	473.2	-7.4%	331.9	601.3
WGT-RF	526609	-24.8%	482.1	-5.6%	336.7	614.5
Panel C: One-stage models						
OLS	562105	-19.7%	497.7	-2.6%	349.1	600.0
LASSO-MIN	552335	-21.1%	485.6	-4.9%	349.3	580.1
RIDGE-MIN	551830	-21.2%	487.7	-4.5%	346.8	589.7
RF	**483993**	-30.9%	**454.9**	-11.0%	331.7	583.0

Table 11: Hospitals, ambulance, doctors in Bratislava: Views

Top 10 important variables via RF model (permutation approach)

Variable	type	Importance	Views for 50% lowest [for 0 if dummy]	Views for 50% highest [for 1 if dummy]	t test [p-value]
Number of words in the title	log	77156.66	943.33	1012.06	0.031
Number of letters in the title	log	64206.75	1023.56	934.88	0.006
Middle school graduate	[1 - yes, 0 - no]	63996.13	821.59	1118.42	0.000
University - master	[1 - yes, 0 - no]	40217.33	1025.89	727.58	0.000
English	[1 - yes, 0 - no]	34012.91	1016.60	707.16	0.000
Slovak	[1 - yes, 0 - no]	27643.15	714.67	1016.20	0.000
Suitable for graduates	[1 - yes, 0 - no]	17924.76	915.00	1042.43	0.000
2 to 3 words in the title	[1 - yes, 0 - no]	14596.24	950.70	1034.38	0.017
1 word in the title	[1 - yes, 0 - no]	9444.94	972.79	1021.67	0.282
4 to 6 words in the title	[1 - yes, 0 - no]	8981.60	960.89	1028.60	0.068

Table 12: Hospitals, ambulance, doctors in Bratislava: Views

Methodological take-away

- One-stage **random forest** has led to most accurate outcomes.
- **Instead** of predicting conversions directly (as now) **one might** predict conversions by dividing predicted reaction with predicted views.
- Highly skewed data - **quantile** based models.
- Stakeholders might be interested in **lower boundaries**: 'With 90% conversion is going to be 1.8% or more.
- Use **interactions**.
- Standardize data handling.

Summary

- Results differ across business area and regions $\rightarrow$ separate models are needed.
- Similar drivers across business area, regions and dependent variables.
- Non-linear models are preferred.
- Accuracy needs to be improved.

Predicting Views, Reactions and Conversions

A Preliminary Analysis via Machine Learning

Patrik Jankovič, Štefan Lyócsa & Miroslav Štefánik

Institute of Economic Research
Slovak Academy of Sciences
Šancova 56, Bratislava, Slovakia

September 24, 2020