Big Data Mining

Influence of Weather Conditions on Energy Consumption in Lisbon and analysis of Domestic and Industrial regions

Professor Nuno Datia, Professora Matilde Pato, Professor Artur Ferreira

NUNO GOMES 18364, RICARDO RAMOS 46638, RAFAEL CARVALHO 47663

Presentation Plan

- Understanding the Data Mining and Contextualization Problem
  - Data Interpretation
- Dataset Characterization and Preprocessing
  - Data Imputation and dimensionality reduction
  - Questioning
- Application of Models and Performance Assessment
  - Model comparison
  - Sampling methods comparison
- Conclusion and future work

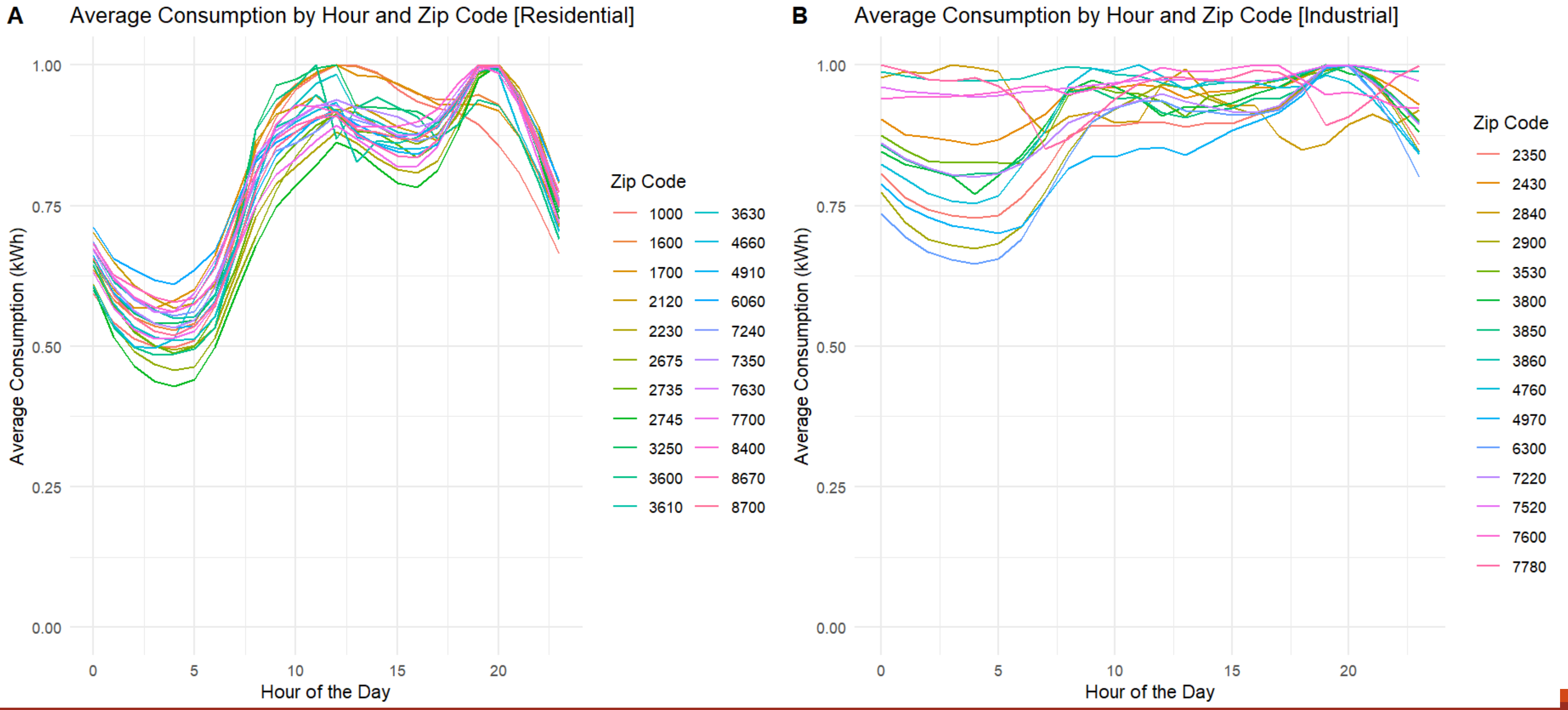
Data Interpretation – Weather

- Weather characteristics in Lisbon in a two-year period
  - 9504 instances with 24 features
  - Imputation based on median, mean, or zero, given standard deviation of original values

name	datetime	temp	feelslike	dew	humidity	precip
0	0	0	0	0	0	0
precipprob	preciptype	snow	snowdepth	windgust	windspeed	winddir
0	744	0	0	0	0	0
sealevelpressure	cloudcover	visibility	solarradiation	solarenergy	uvindex	severerisk
0	0	0	1695	1695	1695	8760
conditions	icon	stations				
0	0	0				

Data Interpretation – Energy

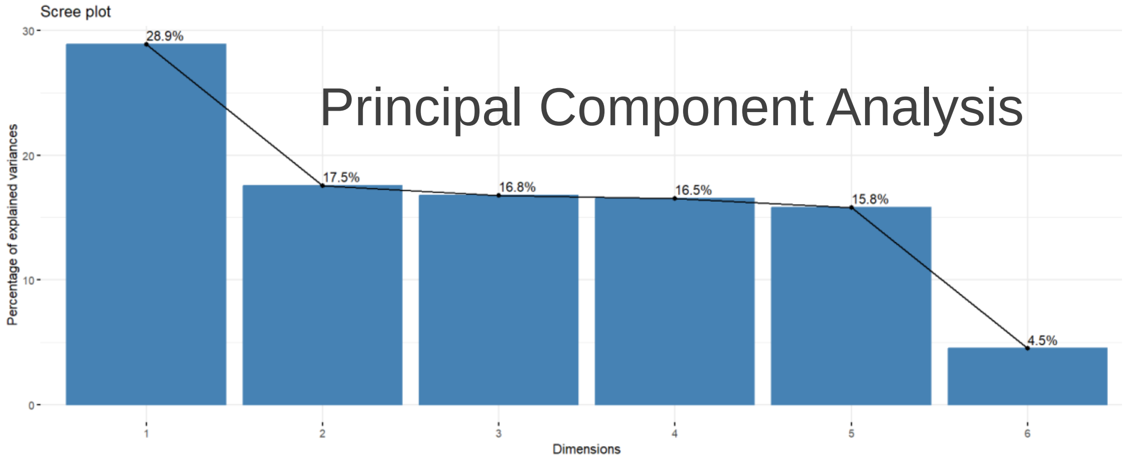
- Energy Consumption for different Zip Codes in Portugal
  - Residential and Industrial categorization (ref: www.pordata.pt)
  - ~3.7M instances with 5 features



Questioning and data pre-processing

- 1 - “What Zip Codes are residential or Industrial given energy consumption characteristics?”
- 2 - “Can we accurately predict energy consumption in Lisbon given expected weather conditions?”
- Feature Selection / Feature Reduction with 95% variance of original data

	Variance Threshold	Fisher's Ratio	Information Gain
visibility	2.312277e-03	60.5000000	0.050868580
humidity	3.742142e-02	37.1306667	0.078975925
Active.Energy..kWh.	2.703371e-02	29.8530769	0.013168736
preciptype	2.621799e-02	28.0900000	0.050505299
temp	2.176463e-02	19.6544444	0.050541986
dew	4.890074e-02	16.1219048	0.060951351



Questioning and data pre-processing

- Data Discretization with Equal Frequency Binning
- Bins then converted into discrete integers
- Labels: 1 – Residential; 2 – Industrial

visibility	humidity	Active.Energy..kWh.	preciptype	temp	dew	feelslike	cloudcover
(-Inf,10]	(95.1,100]	(1.14e+04,1.25e+04]	unknown	(10.2,13]	(10.9,12.2]	(10.2,13]	(89.2,100]
(-Inf,10]	(95.1,100]	(1.05e+04,1.14e+04]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(89.2,100]
(-Inf,10]	(95.1,100]	(9.86e+03,1.05e+04]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(89.2,100]
(-Inf,10]	(95.1,100]	(9.38e+03,9.86e+03]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(89.2,100]
(-Inf,10]	(95.1,100]	(-Inf,9.38e+03]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(89.2,100]
(-Inf,10]	(95.1,100]	(9.38e+03,9.86e+03]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(82.9,89.2]
(-Inf,10]	(95.1,100]	(9.86e+03,1.05e+04]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(82.9,89.2]
(-Inf,10]	(95.1,100]	(1.05e+04,1.14e+04]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(82.9,89.2]
(-Inf,10]	(95.1,100]	(1.25e+04,1.34e+04]	unknown	(-Inf,10.2]	(8.5,10.9]	(-Inf,10.2]	(39.4,51.9]

PC1	PC2	PC3	PC4	PC5	zip_codes	labels
(-1.45e+03,-1.43e+03]	(-Inf,709]	(-Inf,-3.47]	(-Inf,107]	(-Inf,678]	7240	2
(-1.83e+03,-1.78e+03]	(8.91e+03,1.04e+04]	(277,316]	(369,406]	(8.91e+03,1.04e+04]	8700	2
(-1.46e+03,-1.45e+03]	(709,1.1e+03]	(7.78,18.2]	(118,127]	(678,1.07e+03]	2230	2
(-1.93e+03,-1.87e+03]	(1.32e+04,1.48e+04]	(358,405]	(445,489]	(1.32e+04,1.48e+04]	1000	2
(-1.49e+03,-1.47e+03]	(1.45e+03,1.81e+03]	(7.78,18.2]	(118,127]	(1.42e+03,1.78e+03]	3610	2
(-1.78e+03,-1.73e+03]	(8.91e+03,1.04e+04]	(234,277]	(329,369]	(8.91e+03,1.04e+04]	1000	2
(-1.57e+03,-1.53e+03]	(3.19e+03,4.33e+03]	(69,102]	(175,205]	(3.17e+03,4.31e+03]	3600	2
(-1.46e+03,-1.45e+03]	(709,1.1e+03]	(-3.47,7.78]	(107,118]	(678,1.07e+03]	7700	2
(-1.78e+03,-1.73e+03]	(8.91e+03,1.04e+04]	(234,277]	(329,369]	(8.91e+03,1.04e+04]	8700	2
(-1.49e+03,-1.47e+03]	(1.45e+03,1.81e+03]	(-3.47,7.78]	(118,127]	(1.42e+03,1.78e+03]	3610	2
(-1.46e+03,-1.45e+03]	(709,1.1e+03]	(-3.47,7.78]	(107,118]	(678,1.07e+03]	2230	2

Question 1 – Classifying a Zip Code as Industrial or Residential

- Best results for Decision Trees, performance to be investigated
- Logistic Regression displays worst performance overall
- Oversampling gets the best results
- Random seed influences Random Forest and Logistic Regression

	False Positive Rate	Accuracy	Kappa	Pos Pred Value	Neg Pred Value	F1 Score	Area under ROC
Random Forest							
No Sampling	0,23	0,90	0,79	0,78	**0,99**	0,86	0,88
Oversampling	**0,07**	**0,96**	**0,92**	**0,93**	0,98	**0,95**	**0,96**
Undersampling	0,08	**0,96**	**0,92**	0,92	**0,99**	**0,95**	**0,96**
Logistic Regression							
No Sampling	0,29	0,80	0,58	0,71	**0,86**	0,74	0,79
Oversampling	**0,21**	**0,81**	**0,61**	**0,80**	0,82	**0,77**	**0,81**
Undersampling	0,22	0,80	0,60	0,78	0,82	0,76	0,80
Decision Trees							
No Sampling	0,03	0,98	0,95	0,97	0,98	0,97	0,98
Oversampling	**0,02**	**0,99**	**0,98**	**0,99**	**0,99**	**0,99**	**0,99**
Undersampling	**0,02**	**0,99**	**0,98**	**0,99**	**0,99**	**0,99**	**0,99**

CLASS	Instâncias
Industrial	80232
Residencial	117443

Question 2 – Predicting energetic consumption in Lisbon

- Best results for Gradient Boosted Trees
- Simplest model (Linear Regression) is the worst
- Oversampling and undersampling distort the dataset

CLASS	Instâncias
Clear	1189
Overcast	278
Partially Cloudy	3292
Rain, Overcast	10
Rain, Partially Cloudy	52

	Random Forest	Linear Regression	Decision Trees	Gradient Boosted Trees
rmse	1,552	2,48	1,496	**1,092**
mse	2,41	6,151	2,238	**1,191**
r2	0,855	0,63	0,865	**0,928**
mae	1,203	1,959	1,12	**0,757**

Conclusions and Areas for Improvement

- The baseline models perform the worst for both scenarios with oversampling being the best
- Some models were sensitive to the seed value, making them unfit for the problem.
- Add a “Mixed” class for energetic consumption and apply One-vs-Rest model
- Compare performance of reduced datasets vs original datasets
- Further investigate performance of the Decision Trees model in both undersampling and oversampling cases.

Any questions?