Predicția Soldului Total în Sistemul Energetic Național (SEN) pentru Decembrie 2024

Motelică Sandu, Grupa A2

Decembrie 21, 2024

1 Descrierea Problemei

Scopul acestui proiect este de a prezice soldul total al Sistemului Energetic Național (SEN) din România pentru luna decembrie 2024. Datasetul utilizat descrie consumul și producția de energie electrică, segmentate în funcție de surse precum hidro, solar, eolian, cărbune și altele.

Principala provocare constă în adaptarea algoritmilor ID3 și clasificarea bayesiană, proiectați pentru probleme de clasificare, pentru a rezolva o problemă de regresie. De asemenea, datele din decembrie nu pot fi utilizate pentru antrenare, fiind destinate exclusiv testării.

2 Justificarea Abordării

Pentru a rezolva problema regresiei folosind algoritmii menționați, au fost realizate următoarele adaptări:

2.1 Adaptarea Algoritmului ID3

Algoritmul ID3 a fost modificat pentru a se potrivi problemei de regresie prin:

- **Predicția directă a valorilor continue**: Algoritmul a fost adaptat pentru a utiliza *DecisionTreeRegressor*, permițând astfel predicția directă a valorilor continue pentru variabila țintă *Sold[MW]*. Aceasta a reprezentat cea mai performantă metodă.
- **Bucketing pentru variabila țintă**: Într-o altă abordare, valorile soldului au fost împărțite în intervale discrete utilizând binning uniform.

Fiecărui interval i s-a atribuit o valoare reprezentativă, corespunzând mijlocului intervalului. Această metodă a avut rezultate mai slabe din cauza pierderii de informație în procesul de discretizare.

- **Predicția componentelor individuale**: În loc să se prezică direct soldul, algoritmul a fost utilizat pentru a prezice separat *Producția[MW]* și *Consumul[MW]*. Soldul a fost calculat ulterior ca diferența între aceste două valori. Această metodă a avut rezultate variabile, în funcție de calitatea predicțiilor pentru fiecare componentă.
- **Optimizarea hiperparametrilor**: Parametri precum *max_depth*, *min_samples_split* și *min_samples_leaf* au fost optimizați folosind *GridSearchCV*, ceea ce a îmbunătățit performanța modelului și a redus supraînvățarea.

2.2 Adaptarea Clasificării Bayesiene

Clasificarea bayesiană a fost adaptată astfel:

- Toate variabilele continue au fost discretizate utilizând binning uniform.
- Probabilitățile condiționate au fost calculate pentru fiecare combinație de caracteristici, iar predicția finală a fost realizată prin media valorilor țintă corespunzătoare.

2.3 Abordări Multiple de Predicție

Au fost testate mai multe metode:

- Predicția directă a soldului utilizând ID3.
- Predicția componentelor (producție și consum) cu calculul soldului ulterior.
- Predicția soldului utilizând bucătirea valorilor (bucket prediction).
- Regresia bayesiană pe baza variabilelor discretizate.

2.4 Importanța Selectării Caracteristicilor

Un aspect cheie al proiectului a fost ajustarea seturilor de caracteristici utilizate pentru antrenarea modelelor, bazată pe analiza importanței caracteristicilor. În urma utilizării metricei *feature_importances*, s-a observat că variabila *'Year'* avea o importanță zero pentru toate modelele testate. Astfel,

setul inițial de caracteristici *['Year', 'Day_of_Week', 'Hour']* a fost modificat prin eliminarea acestei variabile inutile. Acest proces a permis îmbunătățirea eficienței modelelor și reducerea complexității inutile a datelor.

3 Prezentarea Rezultatelor

Rezultatele obținute pentru fiecare abordare au fost evaluate utilizând metrici standard: Eroare Medie Absolută (MAE), Eroare Pătratică Medie (RMSE) și coeficientul R-squared (R^2). În continuare, este prezentată o analiză detaliată pentru fiecare set de caracteristici utilizat.

3.1 Performanța Abordărilor pe Seturi de Caracteristici

Rezultatele înainte și după optimizarea hiperparametrilor sunt prezentate în Tabelele 1 și 2. Analizăm mai jos impactul acestor optimizări.

Set de Caracteristici	**Metodă**	**MAE**	**RMSE**	**R^2**
'Day_of_Week', 'Hour'	Bayesian	785.37	978.64	-0.17
	Direct Prediction	782.24	977.80	-0.17
	Component Prediction	1005.74	1167.31	-0.66
	Bucket Prediction	717.18	861.50	0.09
'Day_of_Week', 'Hour', 'Consum', 'Intermittent_Production', 'Constant_Production'	Bayesian	458.50	582.50	0.59
	Direct Prediction	393.56	482.16	0.71
	Component Prediction	1412.60	1682.81	-2.45
	Bucket Prediction	389.39	480.74	0.71
'Day_of_Week', 'Hour', 'Consum', 'Productie'	Bayesian	418.48	533.13	0.65
	Direct Prediction	247.86	301.14	0.88
	Component Prediction	735.69	873.30	0.07
	Bucket Prediction	371.80	455.88	0.74
'Consum', 'Productie'	Bayesian	442.32	545.08	0.63
	Direct Prediction	247.86	301.14	0.88
	Component Prediction	292.41	354.03	0.84
	Bucket Prediction	371.80	455.88	0.74

Table 1: Rezultatele metodelor de predicție pentru diferite seturi de caracteristici înainte de adaptarea hiperparametrilor.

Set de Caracteristici	Metodă	MAE	RMSE	R^2
'Day_of_Week', 'Hour'	Bayesian	785.36	978.63	-0.16
	Direct Prediction	776.40	972.57	-0.15
	Component Prediction	1009.33	1174.84	-0.68
	Bucket Prediction	707.34	844.88	0.12
'Day_of_Week', 'Hour', 'Consum', 'Intermittent_Production', 'Constant_Production'	Bayesian	458.50	582.50	0.5859
	Direct Prediction	62.18	83.20	0.99
	Component Prediction	1366.61	1662.72	-2.37
	Bucket Prediction	190.31	227.14	0.93
'Day_of_Week', 'Hour', 'Consum', 'Productie'	Bayesian	418.48	533.13	0.65
	Direct Prediction	**11.31**	**16.64**	**0.99**
	Component Prediction	735.05	885.96	0.04
	Bucket Prediction	173.82	200.30	0.95
'Consum', 'Productie'	Bayesian	442.32	545.08	0.63
	Direct Prediction	9.85	15.55	0.99
	Component Prediction	292.41	354.03	0.84
	Bucket Prediction	173.61	199.94	0.95

Table 2: Rezultatele metodelor de predicție pentru diferite seturi de caracteristici după adaptarea hiperparametrilor

3.1.1 Compararea rezultatelor înainte și după optimizarea hiperparametrilor

După optimizarea hiperparametrilor pentru ID3, s-au observat îmbunătățiri semnificative în performanța modelelor. De exemplu:

- Pentru setul *['Day_of_Week', 'Hour', 'Consum', 'Intermittent Production', 'Constant Production']*, MAE pentru predicția directă a scăzut de la 393.56 la 62.18, iar R^2 a crescut semnificativ la 0.99.
- Pentru setul *['Consum', 'Productie']*, MAE pentru predicția directă a scăzut de la 247.86 la 9.85, iar R^2 a crescut de la 0.88 la 0.99.

3.1.2 Analiza generală a impactului optimizării

Optimizarea hiperparametrilor a contribuit la:

- Reducerea erorilor (MAE și RMSE) pentru majoritatea seturilor de caracteristici, evidențiind o adaptare mai bună a modelelor la date.
- Creșterea semnificativă a coeficientului R^2, ceea ce indică o explicație mai precisă a variației datelor prin modelele folosite.

- O mai bună diferențiere între metodele testate. Predicția directă rămânând cea mai performantă abordare.

3.1.3 Concluzie asupra rezultatelor

Compararea între tabele evidențiază faptul că optimizarea hiperparametrilor pentru ID3 a avut un impact pozitiv clar asupra performanței generale. Seturile de caracteristici mai complexe, cum ar fi *['Day_of_Week', 'Hour', 'Consum', 'Productie']*, au beneficiat cel mai mult de aceste îmbunătățiri.

4 Concluzii

Rezultatele acestui proiect evidențiază că metoda de predicție directă este cea mai eficientă abordare pentru problema predicției soldului energetic. Dintre toate seturile de caracteristici analizate, setul compus din *['Consum', 'Productie']* s-a remarcat prin performanțe excepționale, atingând cel mai scăzut MAE de 9.85 și un coeficient R^2 de 0.99. Aceasta indică faptul că un model simplu, dar bine calibrat, poate genera rezultate remarcabile.

Simplificarea caracteristicilor la un set minimal, precum *['Consum', 'Productie']*, nu doar că a oferit cele mai bune rezultate, dar a demonstrat și o robustețe ridicată în comparație cu seturile mai complexe. Totuși, abordările bazate pe componente au arătat limitări evidente, necesitând îmbunătățiri suplimentare.

În ceea ce privește algoritmii utilizați, adaptarea ID3 și Bayes la probleme de regresie a fost o soluție viabilă, mai ales prin integrarea discretizării variabilelor continue. Succesul acestor metode a fost strâns legat de calitatea preprocesării datelor.

În viitor, există mai multe direcții posibile de îmbunătățire a metodei:

- **Adaptarea parametrului n_bins la discretizare**: Experimentarea cu valori diferite ale numărului de bin-uri pentru variabilele continue ar putea îmbunătăți granularitatea predicțiilor și adaptarea la distribuția datelor.

- **Preprocesări mai eficiente**: O posibilitate este de a prezice individual fiecare componentă (de exemplu, consumul și producția), iar apoi soldul să fie calculat direct ca diferență între acestea. Această abordare ar putea reduce erorile asociate interacțiunilor complexe dintre variabile.

- **Utilizarea mai multor date de antrenament**: Adăugarea mai multor ani de date ar putea îmbunătăți generalizarea modelelor, oferindu-le o înțelegere mai bună a tendințelor pe termen lung.
- **Agregarea datelor după aspecte temporale**: Transformarea datelor de la nivel de minute la nivel orar sau zilnic ar putea simplifica modelele și reduce zgomotul din date, oferind astfel predicții mai stabile.