Tehnike i metode analize podataka

**Primena linearne regresije nad podacima o potrošnji električne energije**

Danilo Milošević 1732

**Opis podataka i problema**

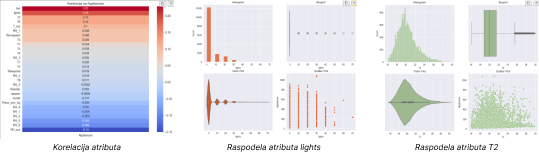
Vrši se analiza podataka o potrošnji električne energije, pri čemu se skup sastoji od 28 atributa i 15788 ulaza. Skup podataka je unapred pripremljen, bez null vrednosti. Pored toga svi atributi su numerički, izuzev atributa date koji je object koji kasnije obrađujemo. Atributi su sledeći:

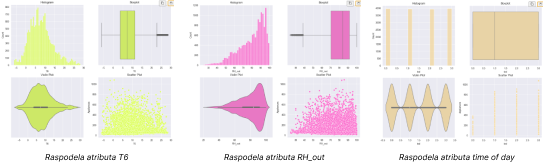
* date - datum i vreme zapisa
* lights - broj sijalica u kući
* T1, RH1 - temperatura i vlažnost vazduha u kuhinji
* T2, RH2 - temperatura i vlažnost vazduha u dnevnoj sobi
* T3, RH3 - temperatura i vlažnost vazduha u sobi za veš
* T4, RH4 - temperatura i vlažnost vazduha u kancelariji
* T5, RH5 - temperatura i vlažnost vazduha u kupatilu
* T6, RH6 - temperatura i vlažnost vazduha van zgrade
* T7, RH7 - temperatura i vlažnost vazduha u sobi za peglanje
* T8, RH8 - temperatura i vlažnost vazduha u dečijoj sobi
* T9, RH9 - temperatura i vlažnost vazduha u sobi za roditelje
* To - temperatura napolju
* Pressure - pritisak
* RH\_out - vlažnost vazduha napolju
* Windspeed - brzina vetra
* Visibility - vidljivost (km)
* Tdewpoint - temperatura potrebna za 100% vlažnost vazduha
* rv1 - nasumična promenljiva
* rv2 - nasumična promenljiva
* **Appliances** - ukupna potrošnja u kWh

Cilj je na osnovu prvih 28 atributa predvideti količinu potrošene energije, odnosno **Appliances** je labela.

**Obrada atributa**

Atribut date modifikujemo, tako da ćemo od datum dobiti doba dana (0-3 - ponoć, jutro, podne, noć). Zatim vršimo analizu korelacije pojedinačnih atributa sa atributom Appliances. Možemo da vidimo da je najveća korelacija sa atributima tod (time of day), lights, T2 i T6 kao i atributom RH\_out. Ispod su prikazane raspodele ovih atributa kao i zavisnosti sa ciljnim atributom.

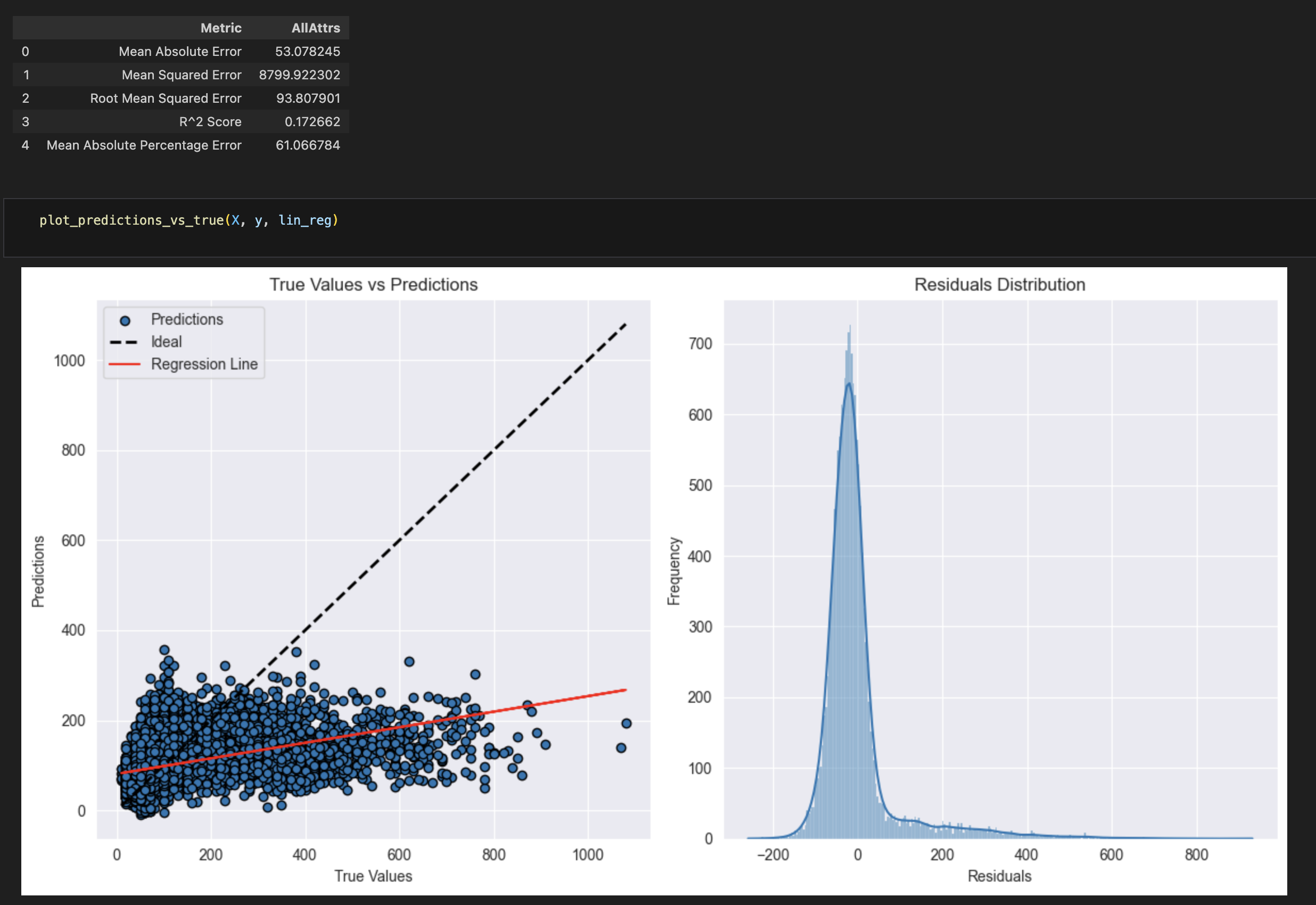




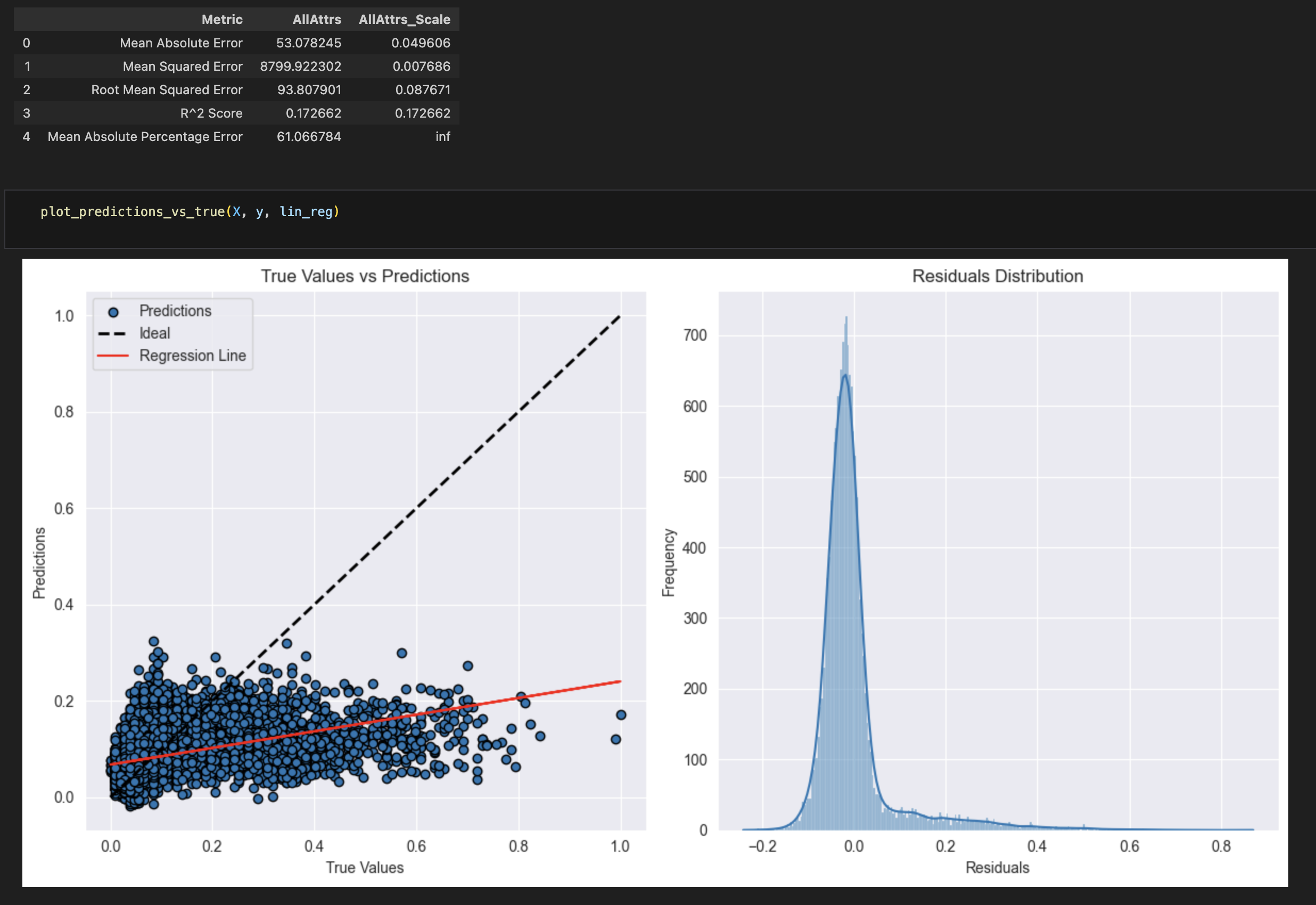
Možemo da primetimo da neki atributi dosta odstupaju od normalne raspodele, zbog čega kasnije tokom treninga isprobavamo da li će normalizacija poboljšati rezultate.

**Trening**

Model linearne regresije je treniran sa 80% početnog skupa podataka, dok je za testiranje korišćeno preostalih 20%. Trenirane su različite kombinacije izbora atributa i izbora transformacija nad podacima. Isprobano je treniranje nad originalnim skupom bez ikakvih transformacija i sa svim atributima. Time se postiže R2 rezultat od 0.173, što se kasnije pokazalo kao jedan od boljih rezultata. Performanse ovog modela su date ispod.



Nakon toga je isproban trening nad skupom podataka koji je prethodno skaliran i treniran nad svim atributima. Rezultati su približno jednaki, izuzev skale naravno.



Zatim je izvršeno treniranje pri čemu su biran najbolji atributi, sa i bez normalizacije kao i treniranje sa svim atributima i normalizacijom. Ipak je najbolje rezultate (po korenu srednje kvadratnoj greške) davao model koji je treniran nad svim atributima koji su prethodno skalirani. Ovo su rezultati koji su ostvareni.



Izvršena je i regularizacija modela, odnosno isproban je i ElasticNet ali rezultati nisu poboljšani.

**Testiranje**

Testiranje modela je dalo blago gore rezultate nego tokom treninga.

