Denisa Dumitrescu

Analiza asupra reusita la loterie

Predictia numarului de castigatori

In acest proiect am decis sa folosesc baza de date a loteriei romane cu gandul ca o sa analizez seturile de numere castigatoare.

Este important de subliniat că rezultatele loteriei sunt aleatorii, iar modelele de predicție nu au nicio valoare reală în acest context și că nu există niciun model care să poată prezice rezultatele loteriei cu exactitate.

Am colectat datele folosite din site-ul arhiva al Lotto si am observat ca dispun de mai multe date precum numarul catigatorilor, datele extragerilor, categorii multiple de castig. In consecinta predictia numerelor castigatoare ar trebuii asociata cu castigul la o categorie, iar pentru a face o predictie ar trebuii sa adaug siruri de numere necastigatoare. Dar pana la etapa aceea mai avem…
In primul rand trebuie explicate datele din excelul meu

Extrageri- Contine numerele de la extrageri 6/49

Test(nefolosite)- Ar fi trebuit sa fie numere de test pentru model

Categoria I ,Categoria II , Categoria III – folosite pentru a analiza numarul de castigatori

Categoria IV- folosit pentru a face o predictie asupra numarului de castigatori, considerata cea mai consistenta categorie

Reduced_data- folosit pentru verificarea predictiei

In toate paginile am pastrat data pentru a putea avea o modalitate de verificare a conexiunii dintre ele.

Predictia Numarului de Castigatori

01
Preprocesarea

02
Antrenarea modelului

03
Concluzii

Preprocesarea

La cititrea datelor, coloanele mele au diferite tipuri de date. Pentru a putea trasa un grafic pe coordonata de timp trebuie sa convertim in datetime. Dar precum se poate observa coloana ‘DATA’ este de forma “Du,27 martie 2021”. Asa ca am facut o functie care taie ziua din fata ‘DATA’ ramand de forma “27 martie 2021”. In continuare am facut un dictionar in care am atribuit fiecarei luna numarul ei corespondent si am inlocuit in ‘DATA’ si am pus puncte intre ele reductand data la o forma convertibila de tipul '%d.%m.%Y'.

Problema a doua intampinata este ca in setul de date colectat marcajul pentru lipsa de catigatori este ‘REPORT’, asa ca pentru categoria 1 si 2 a trebuit sa inlocuiesc cuvantul report cu 0.

O alta problema pe care am sesizat-o tarziu in proiect dar care apartine tot de partea de preprocesare este folosirea ‘.’ in numere. Compilatorul considerea numarul “40.273” ca fiind float64 adica il ia drept 40,273, cand de fapt numarul meu era mult mai mare (40273). Asa ca am eliminat punctele si virgulele si am convertit in int32.

Din acest moment atentia mea era deja deturnata spre analizarea numarului de castigatori din categoria 4 decat a numerelor castigatoare si mi s-a parut o idee interesanta sa incerc sa aplic un algoritm de interpolare pe datele mele, intrucat aveam cateva perioade de aprox 6 luni in care nu aveam date. In mintea mea, interpolarea imi aproxima forma pe care a avut-o graficul anterior in intervalul meu fara date. Dupa mult timp pierdut crezand ca rezultatele mele sunt eronate si ca sigur am facut ceva prost, ajung la ideea ca interpolarea nu face altceva decat sa uneasca punctele intre ele, neschimband cu nimic fundamental graficul de cum il afisa fara a aplica interpolare. (Am decis sa intervalul mare in care lipseau intrari este responsabil pentru lipsa unei concluzii cu privire la incercarea asta). Am redus/izolat datele lasand numai intervalul cel mai complet, restul intrarilor mele ajungand sa le folosesc ca date de test. Avand acum un set de date care puse pe grafic aratau ca un zgomot audio decat ca un set de valori bun pentru o tema, am decisa sa testez si algoritmii de netezire ca sa vedem cum ma pot ajuta, iar cu toate ca Transformata Fourier ne-a fost prezentata ca o alternativa mai buna de netezire, eu am ales sa continui cu media alunecatoare pentru ca rezultatul mi s-a parut mai usor de aproximat la o functie liniara decat rezultatul de la Fourier.

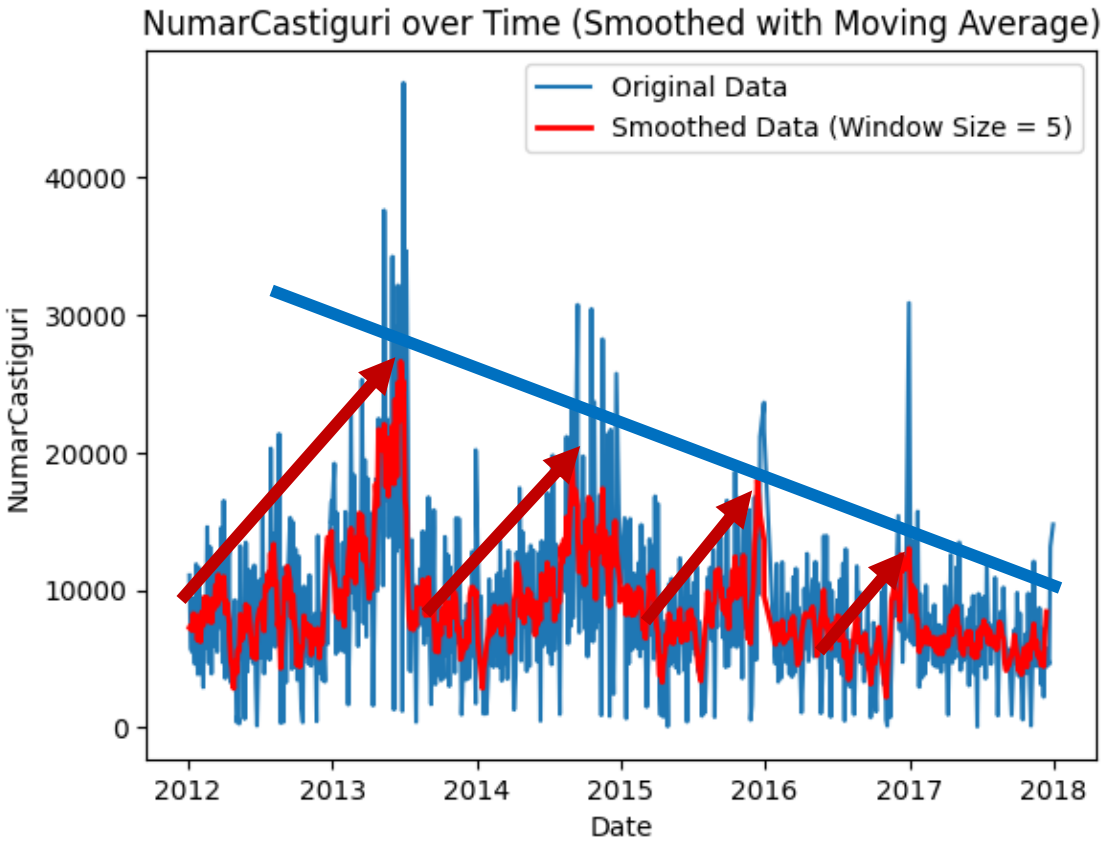
Antrenarea Modelului

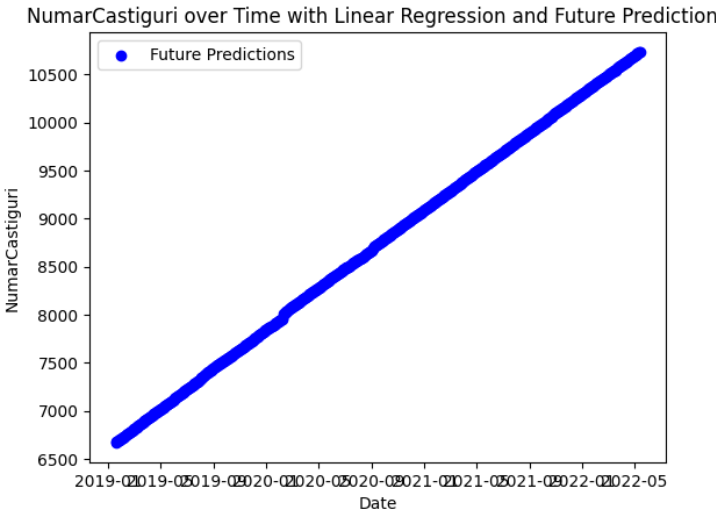
Pentru partea de antrenare a modelului nu am facut nimic mai special fata de laboratorul de regresie prezentat. Am folosit aceeasi tactica, iar la final am comparat cu regresie pe setul de date de test. Am adaugat o parte in care folosesc efectiv modelul de Regresie Liniara pentru prezicere, dar rezultatele au fost departe de realitate.

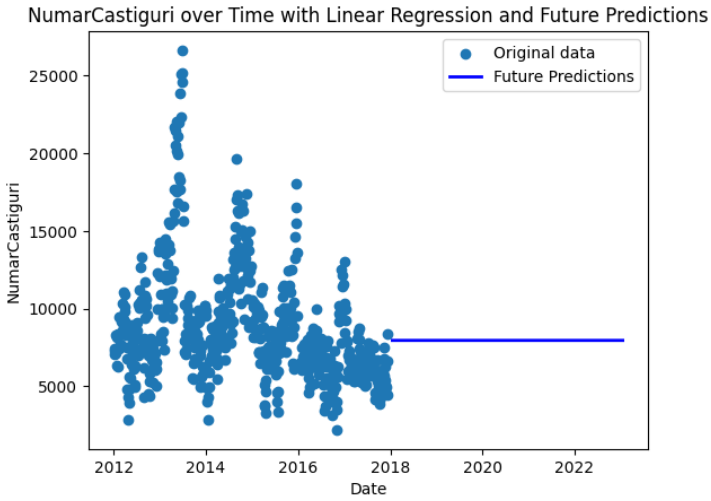
Concluzii

In concluzie, Regresia Liniara nu a fost chiar eficienta in a da un rezultat asupra acestui set de date, estimand o crestere a numarului de catigatori. Acest rezultat nu poate fi etichetat drept incorect caci valorile trec prin perioade de crestere si de descrestere periodic, insa analizand vizual valorile se poate constata ca tendinta numarului de castigatori este de a se micsora. Poate acesta concluzie sa fie afectata de numarul de oameni care participa la lotto?(cu siguranta).

De fapt, concluzia relevanta este ca aceste jocuri sunt imprevizibile si nu exista o reteta matematica pentru a castiga sau a aproxima un numar de castigatori .







In alta ordine de idei, din proiect pana in acest punct a lipsit vreo mentiune asupra numerelor extrase. Acest fapt este datorat esecului pe care l-am avut la partea de lucru cu datele, pentru ca din seturile de numere voiam sa extrag si sa etichez cu 1- castigator ce date aveam de la arhiva loteriei si sa creez eu siruri necastigatoare, urmand sa folosesc RandomForestClassifier din libraria sklearn pentru a face o predictie. Acest lucru mi-a depasit capacitatile caci selectarea a 10 siruri castigatoare combina atat de multe numere incat functia mea nu gasea destule siruri pe care sa le etichetez necastigatoare. Esecul de a construii un set de date echilibrat pentru a putea face o predictie cat de cat aproape de a fi relevant m-a facut sa renunt la partea aceasta atasand proiectului doar niste statistici asupra setului de numere extrase: