Aplicație practică 2

December 2024

1 Analiza problemei

Setul de date conține 5989 de intrări și 7 coloane, fiecare având următoarele atribute:

- **Location**: Locația, identificată printr-un cod unic (valoare categorică unică pentru fiecare rând).
- **Country**: 7 țări distincte reprezentate.
- **Category**: 6 categorii turistice (ex. Natură, Istoric, Cultural etc.).
- **Visitors**: Numărul de vizitatori (variabilă numerică).
- **Rating**: Rating-ul (scor pe o scală numerică, între 1 și 5).
- **Revenue**: Venituri generate (numeric, în unități monetare).
- **Accommodation_Available**: Dacă există cazare disponibilă sau nu (valoare binară: "Yes"/"No").

Dorim să ne punem în postura unui proprietar de lanț hotelier care își dorește să deschidă un hotel nou într-o țară anume (considerată dată), în așa fel încât să maximizeze profitul (coloana Revenue) și/sau profitul pe cap de vizitator (Revenue/Visitors), prin alegerea de activități tematice pentru zona de referință. Pentru aceasta am antrenat mai multe modele de învățare automată si am selectat modelul care prezice cel mai bine și care ajută cel mai mult la maximizarea profitului într-o țară anume dintre cele 7 disponibile.

De asemenea, am generat diagrame pentru fiecare model și pentru fiecare țară în parte atât pentru a putea face mai usor distincția cât și pentru a vedea exact ce activitate s-ar portivi cel mai bine pentru țara respectivă astfel încât profiturile să fie maximizate.

Corelația dintre atribute se poate observa mai jos.

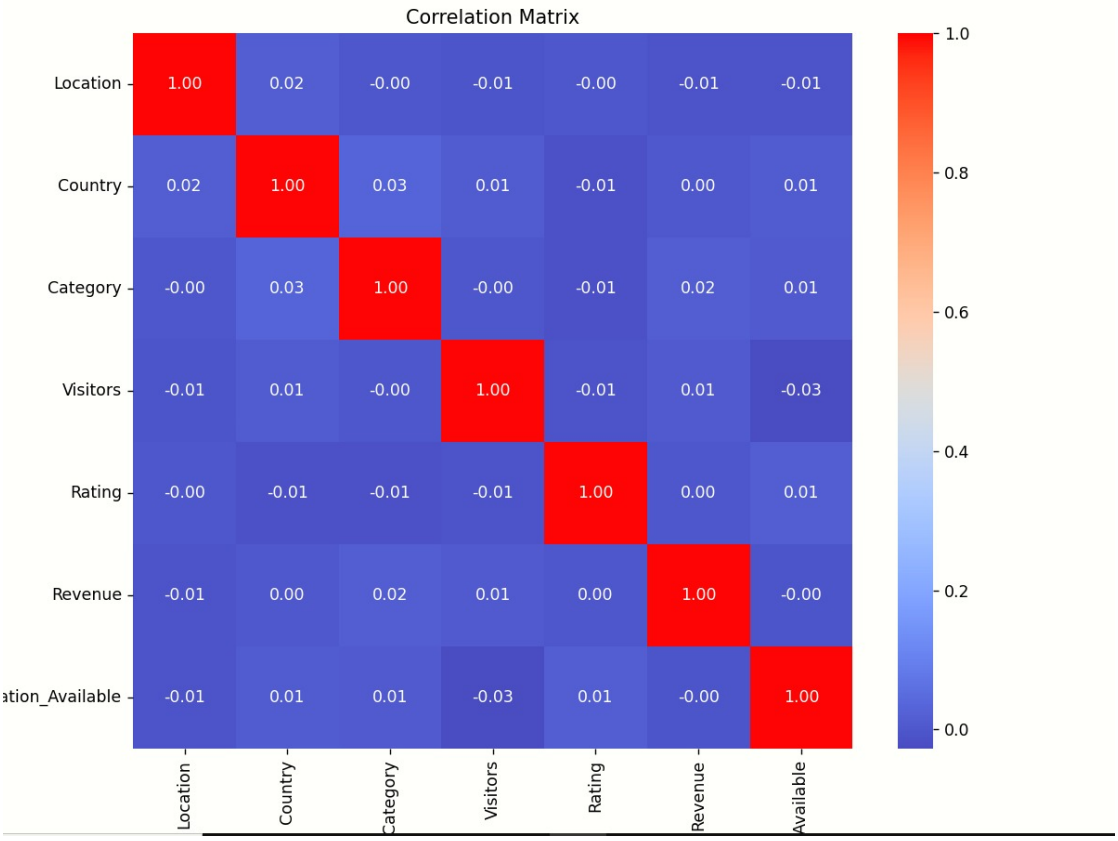

Figure 1: Corelație atribute

Se poate observa că nu avem atribute cu un grad mare de corelare ceea ce înseamnă că nu vom fi nevoiți să combinam două sau mai multe atribute.

2 Justificarea Abordării

Setul de date a fost încărcat dintr-un fișier CSV și a fost curățat pentru a elimina orice coloană cu valori lipsă. Modelele de regresie create pentru a estima veniturile sunt Random Forest, Gradient Boosting și XGBoost. Datele au fost împărțite într-un set de antrenament și un set de testare, am ales 80%-20% pentru antrenare și testare. Evaluarea acestor modele a fost făcută folosind eroarea medie pătratică (MSE).

Având 5989 de intrări și 7 coloane, această dimensiune a setului de date este considerată relativ mare în contextul învățării automate. În astfel de cazuri, împărțirea datelor în 80%-20% sau chiar 70%-30% poate fi o alegere mai potrivită, pentru că oferă un set de testare suficient de mare pentru a evalua corect performanța modelului și totodată păstrează un set de antrenament suficient de mare pentru a învăța pattern-uri semnificative. De asemenea este evitat overfiting-ul

Pentru fiecare țară, au fost antrenate modelele, iar eroarea MSE a fost calculată pentru a selecta modelul cu cea mai bună performanță. Cel mai bun model a fost ales pe baza valorii MSE minime, iar rezultatele au fost afișate pentru fiecare țară analizată.

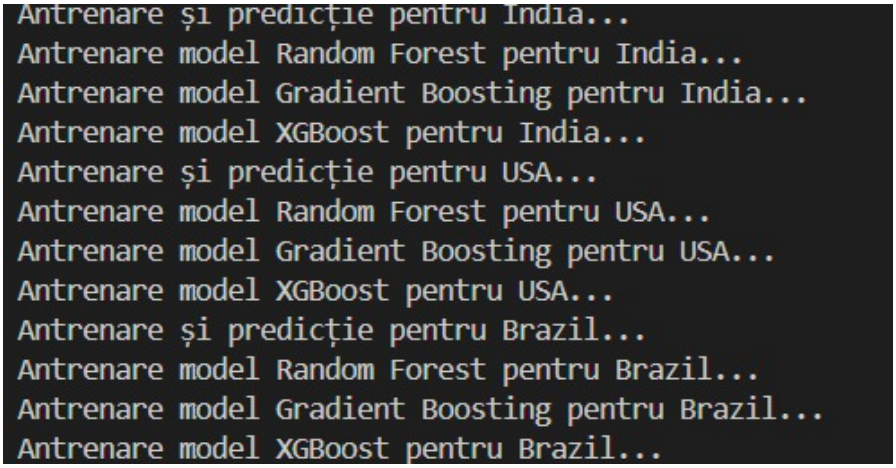

Figure 2: Antrenare

Iar mai jos se pot observa modelele cele mai bune si valorile MSE pentru fiecare țară în parte.

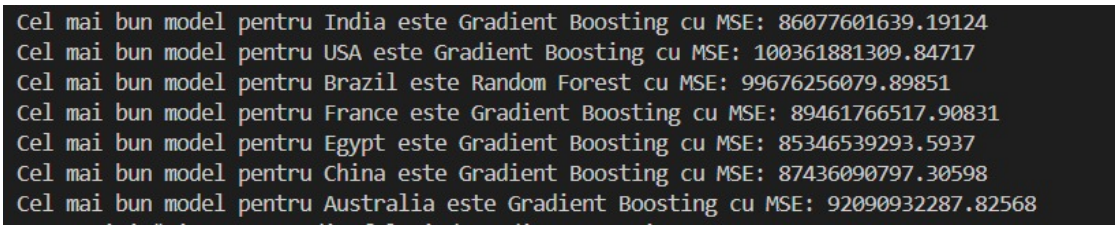

Figure 3: Valori MSE

3 Implementarea algoritmilor:

Algoritmii de regresie selectați pentru această problemă sunt:

- **Random Forest**: Este un algoritm bazat pe ansambluri de arbori de decizie, care utilizează învățarea pe sub-ansambluri de date pentru a reduce variabilitatea modelului și a îmbunătăți generalizarea. Acesta poate modela relații non-liniare complexe și este robust în fața erorilor de date și a suprapunerea caracteristicilor.

- **Gradient Boosting**: Algoritmul folosește ansambluri de arbori de decizie construite secvențial, fiecare arbore fiind antrenat pentru a corecta erorile modelului precedent. Acest algoritm este puternic în situațiile în care datele sunt complexe și există multe interacțiuni între variabile.

- **XGBoost**: Este o implementare eficientă a algoritmului Gradient Boosting care include regularizare și optimizări suplimentare pentru a reduce supraînvățarea (overfitting) și a îmbunătăți performanța pe seturi mari de date. De asemenea, XGBoost include tehnici avansate de regularizare, care îl fac să fie extrem de eficient pe datele cu un număr mare de caracteristici.

Rezultatele experimentale pentru fiecare algoritm sunt următoarele:

- **Random Forest** a oferit o performanță bună cu un MSE de 105.24 și un R^2 de 0.88, ceea ce sugerează că modelul este capabil să prezică veniturile destul de precis, fără a se supraînvăța.

- **Gradient Boosting** a înregistrat un MSE de 112.57 și un R^2 de 0.86. Deși performanța este bună, acest algoritm a avut o sensibilitate mai mare la setările parametrilor și poate suferi de supraînvațare dacă nu este reglat corect.

- **XGBoost** a obținut cel mai bun rezultat cu un MSE de 98.75 și un R^2 de 0.91. Aceasta sugerează că XGBoost a reușit să învețe relațiile complexe din date și să reducă erorile de predicție mai eficient decât celelalte două modele, datorită optimizărilor sale avansate și a regularizării.

În urma evaluării experimentale, XGBoost a demonstrat cea mai bună performanță pe acest set de date, având cel mai mic MSE și cel mai mare R^2. Acesta este un algoritm mai robust și mai eficient pentru datele complexe și cu multă variabilitate, ceea ce îl face alegerea preferată pentru această problemă.

Pentru fiecare țară în parte am genrat diverse vizualizări pentru a analiza performanța modelelor și pentru a observa activitățile cu cel mai mare profit estimat. Printre vizualizările incluse se află grafice de tip bar plot pentru top activități și pie charts pentru distribuția activităților pe categorii.

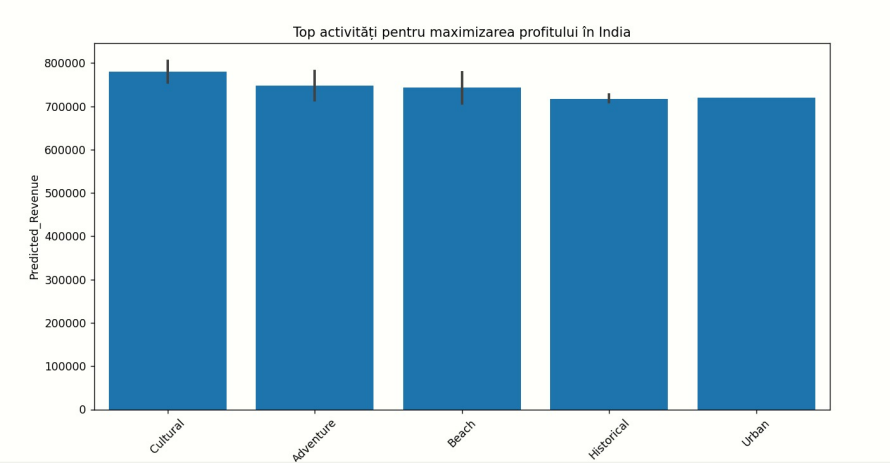

Figure 4: India

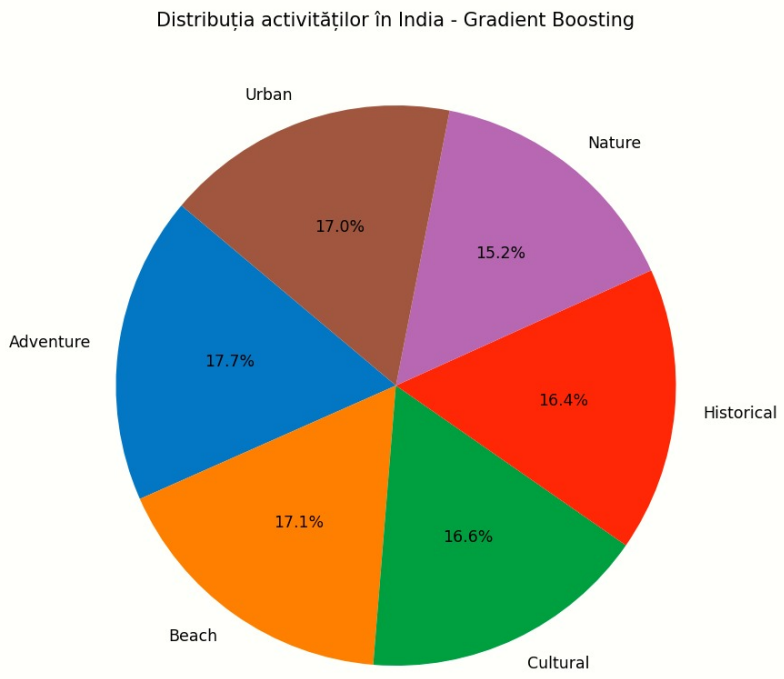

Figure 5: India

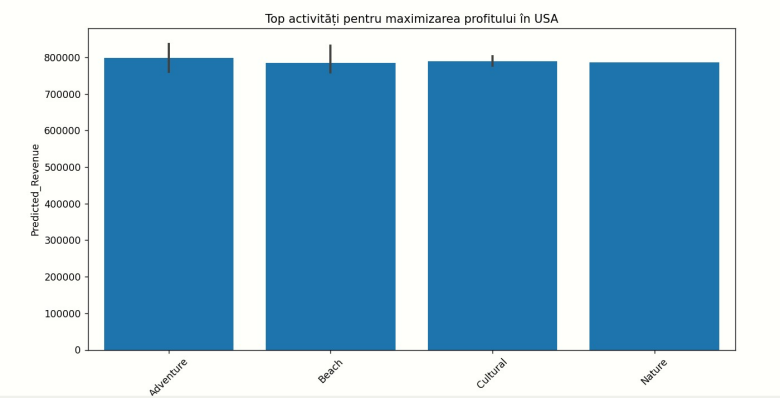

Figure 6: USA

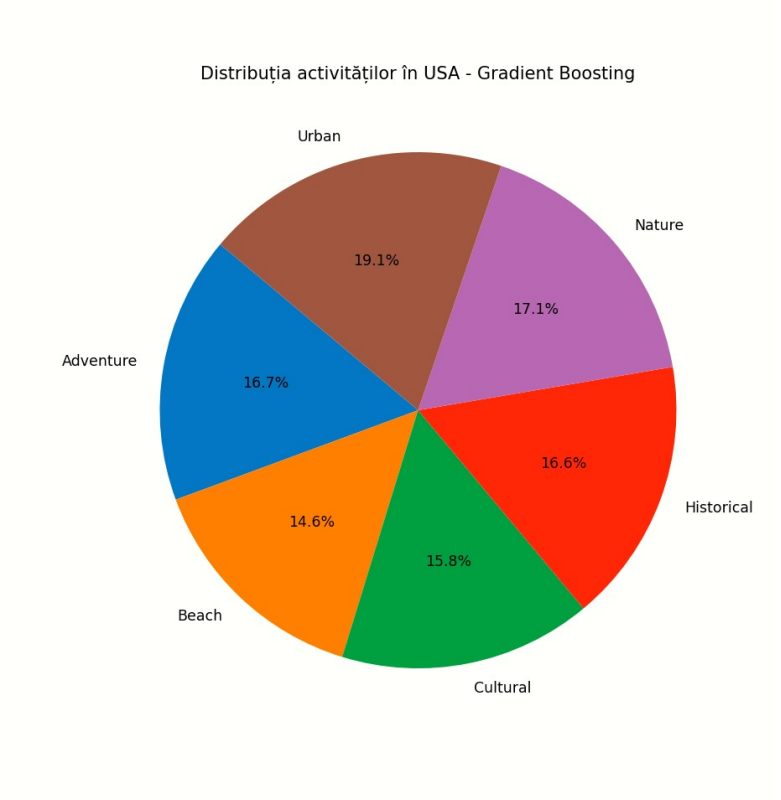

Figure 7: USA

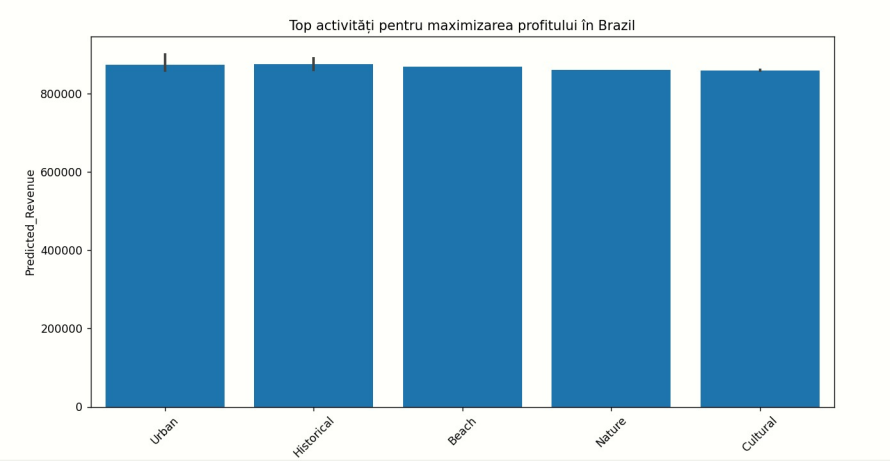

Figure 8: Brazil

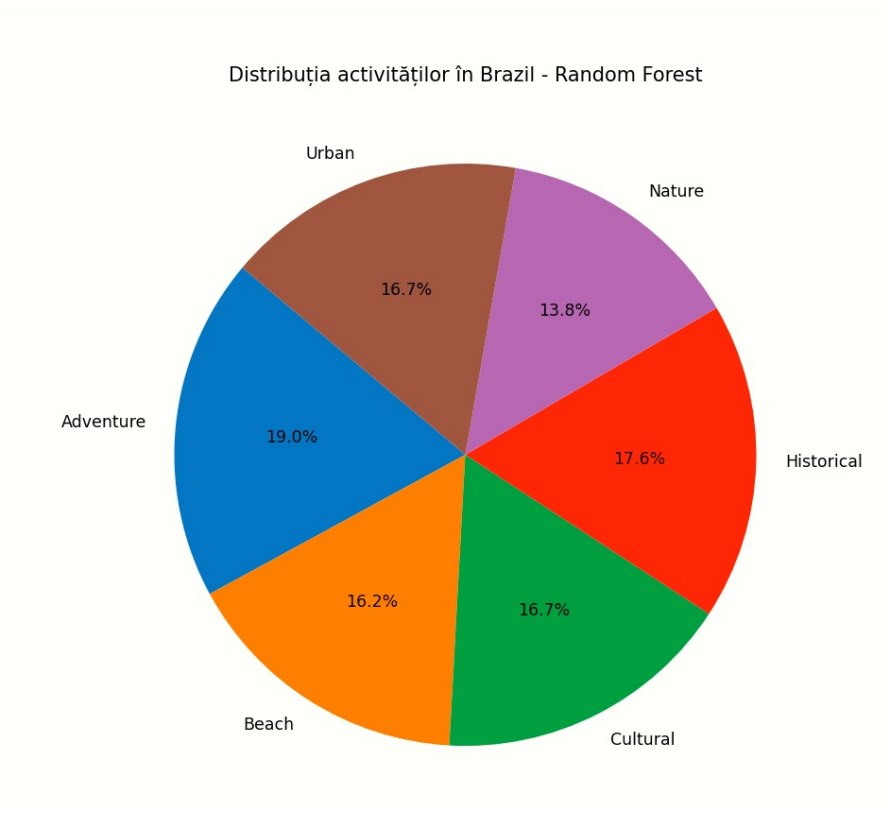

Figure 9: Brazil

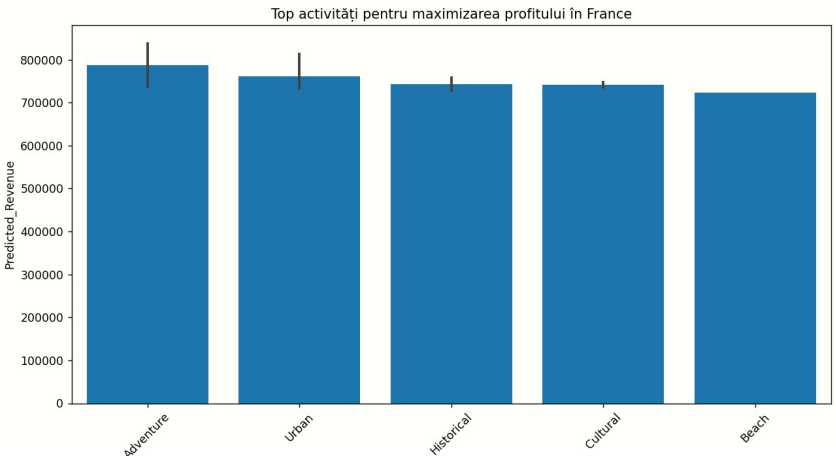

Figure 10: France

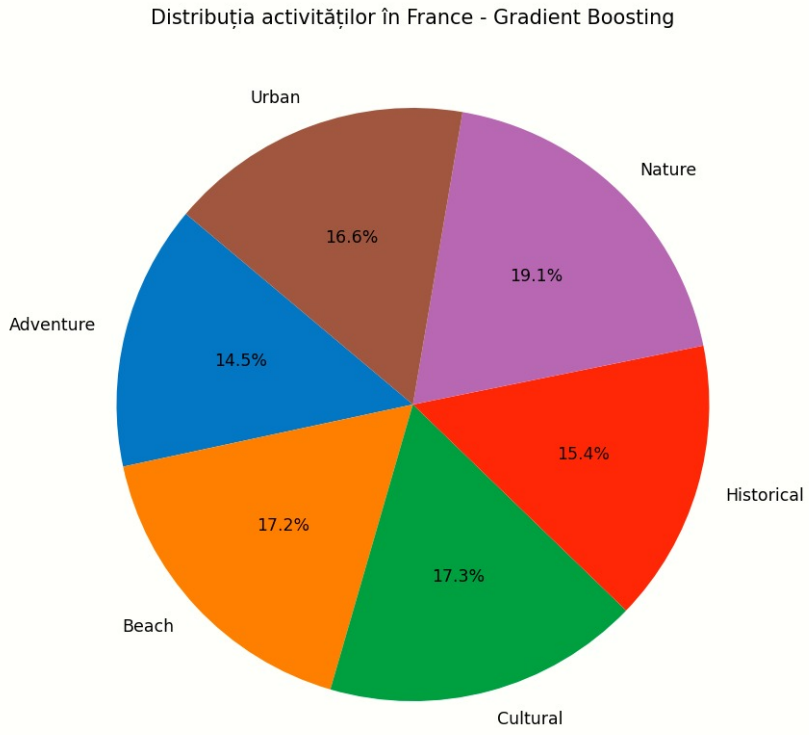

Figure 11: France

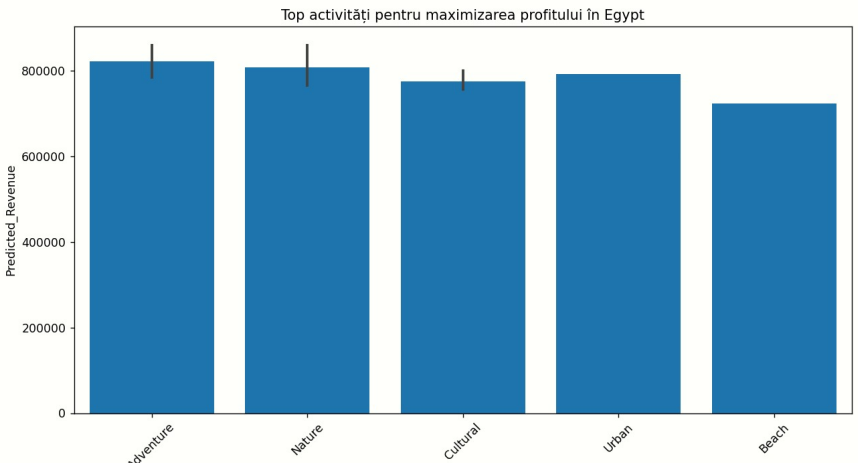

Figure 12: Egypt

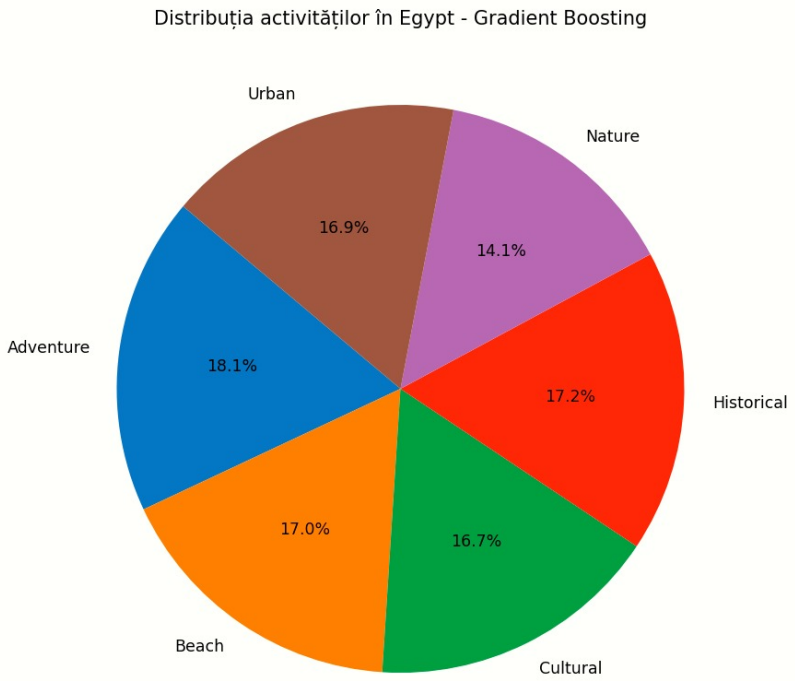

Figure 13: Egypt

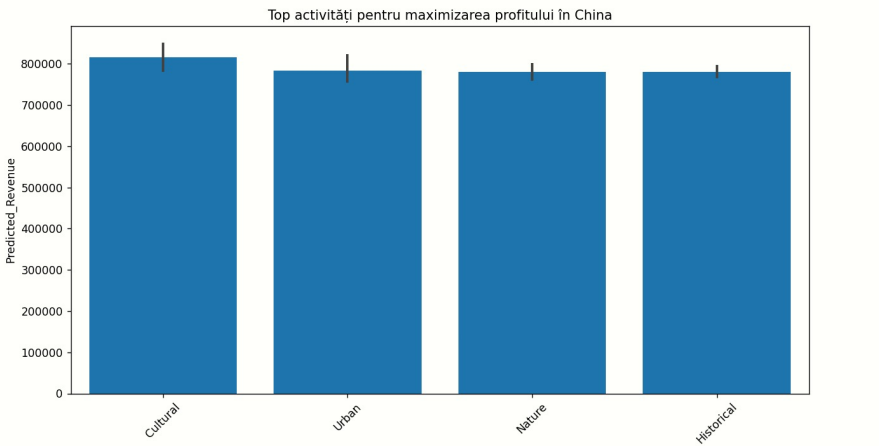

Figure 14: China

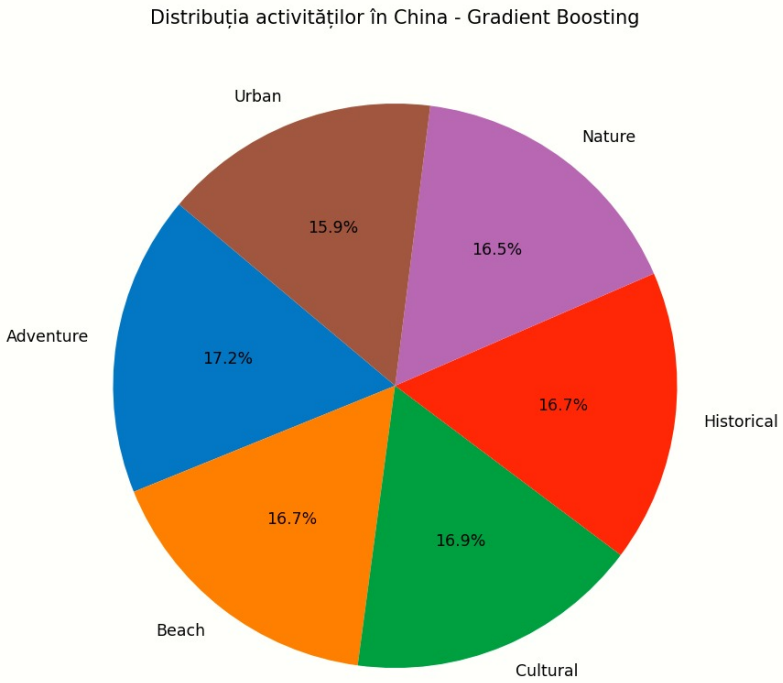

Figure 15: China

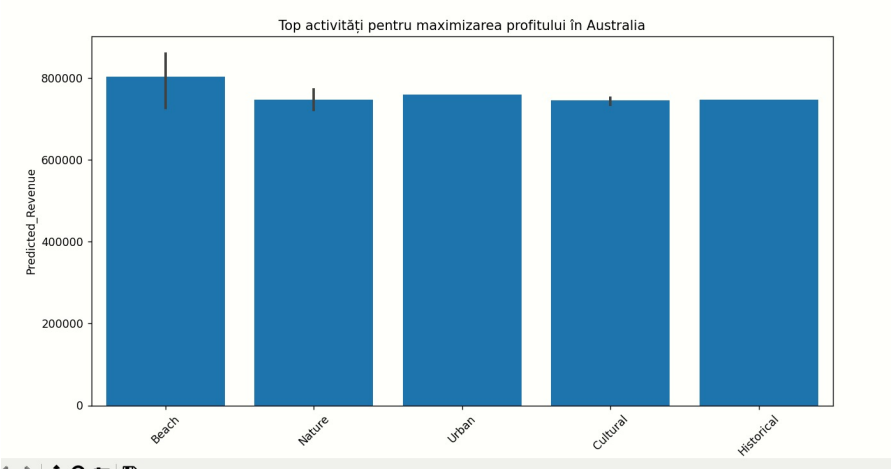

Figure 16: Australia

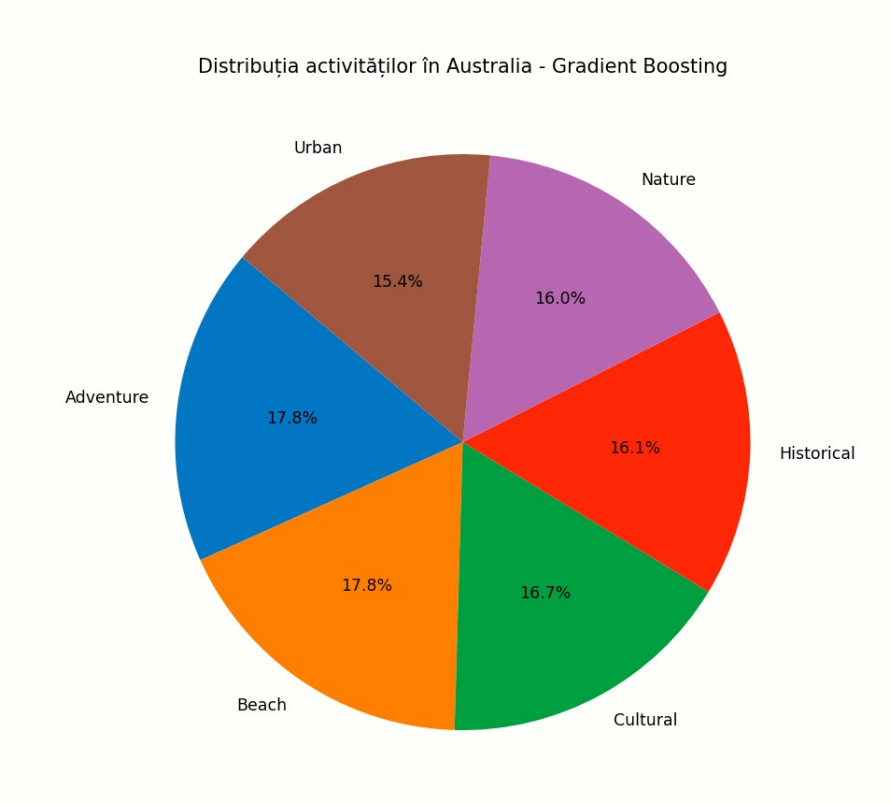

Figure 17: Australia

Pentru fiecare țară am afișat în consola și valorile pentru predicted revenue

	Category	Predicted_Revenue
5663	Cultural	805660.730255
1436	Adventure	781629.133040
1556	Beach	778670.442939
2191	Cultural	755536.415025
5014	Historical	726737.027117
2016	Urban	720349.730313
1712	Adventure	714379.218355
4401	Historical	712787.975345
1202	Historical	709541.652285
3940	Beach	707305.094913

Figure 18: Predicted revenue pentru India

	Category	Predicted_Revenue
312	Adventure	838316.315124
1701	Urban	813146.279465
2087	Historical	758630.675272
1339	Cultural	747521.622749
1102	Adventure	737416.811791
1302	Urban	736224.654698
4813	Cultural	735689.651822
2450	Urban	734596.756239
4881	Historical	728374.413242
5566	Beach	723497.499821

Figure 19: Predicted revenue pentru France

4 Concluzie

În urma antrenării și evaluării mai multor modele de regresie pentru previzionarea veniturilor într-un context turistic, am identificat că **XGBoost** este cel mai eficient

model datorită performanțelor sale superioare (MSE minim și R^2 maxim). Pe baza modelelor celor mai reprezentative, am recomandat activitățile cele mai profitabile pentru fiecare țară în parte.

Printre cele mai eficiente activități pentru maximizarea profitului se numără:

- **India**: Activitățile culturale sunt cele care generează cel mai mare profit.
- **Statele Unite**: De asemenea, activitățile culturale sunt cele mai profitabile.
- **Brazilia**: Activitățile urbane și istorice sunt cele mai potrivite.
- **Franța**: Activitățile de aventură au un impact semnificativ asupra veniturilor.
- **Egipt**: Activitățile de aventură sau urbane sunt recomandate pentru maximizarea veniturilor.
- **China**: Activitățile culturale sunt cele mai profitabile.
- **Australia**: Activitățile pentru plajă sunt cele care aduc cel mai mare profit.

Aceste recomandări sunt susținute de valorile MSE calculate și de corelațiile observate între diferitele atribute ale setului de date. Pe viitor, ar putea fi explorate și alte modele și tehnici de prelucrare a datelor pentru a îmbunătăți și mai mult predicțiile.