Predikcija plata u ekonomski razvijenim državama

Katarina Perović E2 131/2024, Milica Petrović E2 124/2024, Ana Radovanović E2 158/2024

1. Definicija problema

Potrebno je analizirati podatke o platama koji obuhvataju informacije o zaposlenima iz SAD-a, Kine, Velike Britanije, Kanade i Australije, i razviti modele koji će omogućiti preciznu predikciju plata na osnovu godina starosti, pola, nivoa obrazovanja, naziva pozicije i godina iskustva.

2. Motivacija problema rešavanog u projektu

Precizne procene plata bi pomogle poslodavcima da bolje razumeju očekivanja tržišta, definišu konkurentne ponude za zapošljavanje i optimizuju troškove, dok bi kandidatima pružile jasniji uvid u raspon plata za njihove veštine i iskustvo. Ovaj rad bi mogao da doprinese većoj transparentnosti i efikasnosti na tržištu rada.

3. Relevantna literatura

3.1 Salary Prediction using Random Forest with Fundamental Features [pdf]

Tema rada: Rad se bavi predikcijom godišnjeg prihoda uz pomoć Random Forest modela, gde se prihod pojedinca procenjuje na osnovu demografskih, obrazovnih i profesionalnih podataka. Pored toga, rad se fokusira i na poređenje efikasnosti Random Forest modela sa drugim modelima: Logistic Regression, KNN, Naïve Bayes i Decision Tree nad istim skupom podataka.

Podaci: Rad koristi podatke iz skupa Adult iz UCI Machine Learning repozitorijuma, sa informacijama iz popisa stanovništva SAD-a iz 1994. godine. Skup sadrži 32.561 zapis sa 15 atributa, uključujući numeričke (starost, broj godina obrazovanja, kapitalni dobitak, kapitalni gubitak i radni sati nedeljno) i kategorijalne (vrsta posla, težinski faktor, obrazovni nivo, bračni status, zanimanje, porodični status, rasa, pol i zemlja porekla).

Metodologija: Korišćen je Random Forest, metoda koja kombinuje stabla odlučivanja, uz bootstrap uzorkovanje za generisanje podskupova podataka i nasumični izbor atributa za svaki čvor. Kategorijalni podaci iz skupa Adult transformisani su u numeričke pomoću One-Hot Encoding-a. Izlaz iz modela bio je godišnji prihod (≤50k ili >50k). Finalna predikcija dobijena je kombinovanjem rezultata svih stabala, a performanse su upoređene sa metodama: Decision Tree, Logistic Regression, KNN i Naïve Bayes.

Evaluacija rešenja: Za evaluaciju modela korišćeni su metrički pokazatelji tačnost (Accuracy) i AUC (Area Under the Curve), koji mere sposobnost modela da pravilno klasifikuje primere i razlikuje klase. Random Forest je postigao najbolje rezultate sa tačnošću od 92.5% i AUC-om od 0.894, dok su ostali modeli imali niže performanse: Logistic Regression (76.0%, 0.831), KNN (91.6%, 0.877), Naïve Bayes (79.2%, 0.840) i Decision Tree (76.4%, 0.861).

Zaključak: U radu planiramo da iskoristimo Random Forest algoritam, ali za razliku od pristupa opisanog u ovom radu, fokusiraćemo se na regresorski problem, a ne na klasifikaciju. Koristićemo i KNN i Decision Tree metode. Takođe, razlike će biti u tome što će naš skup podataka imati drugačije atribute.

3.2 Employee Salary Prediction in HRMS Using Regression Models [pdf]

Tema rada: Predikcija plata zaposlenih radi povećanja tačnosti, pravičnosti i transparentnosti u određivanju plata pomoću Random Forest Regressor, Gradient Boosting Regressor i LGBM Regressor modela, koristeći HRMS podatke koji sadrže informacije o obrazovanju, godinama iskustva, radnim pozicijama, nivoima zaposlenja, sektoru rada, mesečnim primanjima i drugim relevantnim faktorima.

Podaci: Dataset korišćen u istraživanju preuzet je sa Kaggle pod nazivom "IBM HR data for Performance." Sadrži 1470 redova i 35 kolona, obuhvatajući atribute, kao što su: godine rada, mesečna primanja, radni sati, obrazovni nivo, sektor zaposlenja, pol itd.

Metodologije: Uključivala je pripremu podataka (uklanjanje duplikata, konverziju kategorijalnih u numeričke, uklanjanje outliera i standardizaciju). Korišćeni su modeli Random Forest Regressor, Gradient Boosting Regressor i Light Gradient Boosting Machine Regressor, uz hiperparametarska podešavanja i primenu metode .fit. Izlaz iz modela bila su mesečna primanja.

Evaluacija rešenja: Korišćeni su koeficijent determinacije R^2 i tačnost. Dataset je podeljen na trening (70%) i test (30%) skupove. Random Forest je postigao 96% tačnosti (R^2 = 0.80), Gradient Boosting 95% (R^2 = 0.79), a LGBM Regressor 99% (R^2 = 0.77), čime se LGBM izdvojio kao najprecizniji model.

Zaključak: U radu planiramo da iskoristimo Random Forest Regressor algoritam jer je pokazao visoku tačnost u ovoj oblasti, kao i metodu evaluacije R^2. Razlike će biti u tome što će naš skup podataka imati drugačije atribute i za poređenje performansi modela koristićemo druge modele, u odnosu na one koji su ovde dati u radu.

3.3 Salary Prediction Model for IT Professionals in Ukraine [pdf]

Tema rada: Tema je razvoj modela za predikciju plata IT profesionalaca u Ukrajini pomoću Ordinary Least Squares, Linear Regression, Random Forest i Backpropagation Neural Networks algoritama, koristeći demografske i profesionalne faktore iz anketa IT industrije Ukrajine.

Podaci: Podaci za istraživanje prikupljeni su putem anketa DOU zajednice, najveće IT zajednice u Ukrajini, u periodu 2020–2023. Originalni dataset imao je 77.094 zapisa, ali je nakon čišćenja sveden na 62.890 redova sa 174 atributa. Korišćeni atributi uključuju godine, lokaciju, pol, obrazovanje, iskustvo, poziciju, nivo znanja engleskog jezika itd.

Metodologije: U preprocesiranju su uklonjeni outliers, tretirani nedostajući podaci, konsolidovane varijable i plata je logaritamski transformisana radi bolje raspodele. Ordinary Least Squares je korišćen za procenu značajnih promenljivih (p-vrednosti < 0.05) uz MinMax skaliranje (0–1) i uklanjanje multikolinearnosti. Linear Regression je treniran na ovim promenljivama uz uklanjanje outliera. Random Forest je koristio Friedman MSE za minimizaciju varijanse i stabilnost kroz slučajne podskupove. Backpropagation Neural Networks su primenile sigmoid funkciju, Adam optimizator i Dropout regularizaciju uz normalizaciju varijabli (0–1). Izlaz za sve modele bila je logaritamski transformisana plata.

Evaluacija rešenja: Evaluacija je sprovedena korišćenjem R^2, MAE, MAPE i RMSE metrika kroz stratifikovanu 4-fold cross-validaciju i testiranje na nezavisnom skupu. Tokom validacije, neuronske mreže su postigle R^2 = 75.89%, Random Forest R^2 = 75.81%, a Linear Regression R^2 = 72.13%. Na test skupu, Random Forest je imao najbolji rezultat sa R^2 = 77.30%, dok su neuronske mreže postigle R^2 = 77.11%, a Linear Regression R^2 = 72.79%.

Zaključak: U našem radu koristićemo Random Forest Regressor i Linear Regression algoritme, kao i metode evaluacije R^2, MAE i MAPE. Razlike će biti u tome što će naš skup podataka imati drugačije atribute i performanse će se porediti sa nekim drugim modelima u odnosu na ove.

4. Skup podataka

Podaci koji će biti korišćeni su iz Kaggle skupa Salary by Job Title and Country. Dati skup podataka ima 6684 reda i sadrži sledeće atribute: Age (godine starosti zaposlenog), Gender (pol zaposlenog), Education Level (nivo obrazovanja, gde je 0: High School, 1: Bachelor Degree, 2: Master Degree, 3: PhD), Job Title (naziv pozicije), Years of Experience (godine radnog iskustva), Salary (godišnja zarada izražena u američkim dolarima), Country (zemlja zaposlenja, sa mogućim vrednostima: USA, China, UK, Canada, i Australia), Race (rasa zaposlenog) i Senior (indikator da li je zaposleni na višoj poziciji).

5. Metodologija

Nakon analize podataka, sprovešćemo pretprocesiranje koje uključuje čišćenje podataka, primenu Label Encoding-a za kategorizovane podatke poput pola, kao i One-Hot Encoding-a za kategorije s više vrednosti, poput naziva poslova. Ulazi u modele biće sledeći atributi: Age (godine starosti), Gender (pol), Education Level (nivo obrazovanja), Job Title (naziv pozicije) i Years of Experience (godine radnog iskustva). Kao izlaz modela predviđaćemo godišnju platu izraženu u američkim dolarima. Za modelovanje ćemo koristiti algoritme: Random Forest Regressor, Linear Regressor, K-Nearest Neighbors (KNN) i Decision Tree Regressor. Nakon treninga i testiranja modela, vršićemo njihovu evaluaciju i poređenje.

6. Metod evaluacije

Za evaluaciju predikcije plata koristiće se metrike MAE, MAPE, MSE i R^2, koje mere odstupanje između predviđenih i stvarnih vrednosti plata. Podaci će biti podeljeni na trening i test skup u razmeri 80:20.