Identifikacija najvažnijih značajki u modelu predviđanja uspješnosti poduzeća

Kuliš, Ivana

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 708

IDENTIFIKACIJA NAJVAŽNIJIH ZNAČAJKI U MODELU PREDVIĐANJA USPJEŠNOSTI PODUZEĆA

Ivana Kuliš

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 708

IDENTIFIKACIJA NAJVAŽNIJIH ZNAČAJKI U MODELU PREDVIĐANJA USPJEŠNOSTI PODUZEĆA

Ivana Kuliš

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 30. rujna 2024.

DIPLOMSKI ZADATAK br. 708

Pristupnica:	**Ivana Kuliš (0036525037)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentorica:	doc. dr. sc. Jelena Božek

Zadatak: **Identifikacija najvažnijih značajki u modelu predviđanja uspješnosti poduzeća**

Opis zadatka:

Uspješnost poduzeća na međunarodnoj razini izražava se rangom koji u obzir uzima prihode. Međutim, na uspješnost poduzeća odnosno prihode utječu i druge značajke poput broja zaposlenika te pripadajućeg sektora poslovanja. Korištenjem računalnih modela i većeg broja značajki moguće je predvidjeti uspješnost poduzeća. Identifikacijom značajki koje su najvažnije za predviđanje moguće je u poslovanju utjecati na identificirane značajke te time poboljšati uspješnost poduzeća. U ovom radu potrebno je opisati značajke korištene u javno dostupnim podacima Fortune 1000 koji sadrže tisuću najuspješnijih poduzeća u SAD-u rangiranih po prihodu. Opisati modele strojnog učenja za predviđanje uspješnosti poduzeća te metode za identifikaciju najvažnijih značajki. Cilj rada je implementirati model za predviđanje promjene u rangu uspješnosti poduzeća korištenjem javno dostupnih podataka Fortune 1000. Istražiti kako sektor ili industrija u kojoj poduzeće posluje utječe na njegovu uspješnost s obzirom na financijske performanse. Korištenjem modela istražiti najvažnije značajke koje utječu na predviđanje tržišne vrijednosti poduzeća i prihoda. Usporediti dobivene rezultate s postojećim istraživanjima.

Rok za predaju rada: 14. veljače 2025.

Zahvaljujem profesorima FER-a na znanju koje su mi prenijeli tijekom studija, a naročito mojoj mentorici, doc. dr. sc. Jeleni Božek na svoj pomoći i uloženom vremenu. Zahvaljujem kolegama s faksa, koji su učinili moje studiranje malo lakšim, ali zato neizmjerno ljepšim. Posebno hvala mojim prijateljima koji su me ohrabrivali i vjerovali u mene čak i kad sama nisam. Hvala mojoj sestri što je uvijek navijala za mene i s ponosom pričala o svojoj maloj sestri "inženjerki". Neizmjerno hvala mojim roditeljima koji su mi pružili sve, a na meni je bilo samo da učim. A za kraj, hvala mlađoj Ivani koji je došla iz malog grada, ali je uvijek imala velike snove.

Sadržaj

1. Uvod

Strojno učenje postaje sve češće korišten alat u istraživanjima vezanim uz poduzetništvo, financije i ulaganja. Primjenom algoritama strojnog učenja moguće je doći do važnih uvida te identificirati strategije za unapređenje poslovanja i financijskih rezultata poduzeća. Uspjeh poduzeća može se definirati na različite načine, a u ovom radu će fokus biti na rangu na ljestvici Fortune 1000, koja rangira američka poduzeća prema prihodima.

Dosad je provedena nekolicina istraživanja koja također koriste strojno učenje za predviđanje poslovnog uspjeha poduzeća. Gao, Luo i Pan su u svom radu [1] koristili podatke s Crunchbasea za klasifikaciju start-upova kao uspješnih (prošli su M&A ili postali javni) i neuspješnih. Autori uspoređuju tri algoritma: logističku regresiju, slučajnu šumu i K-najbližih susjeda (KNN), pri čemu je KNN pokazao najbolju F1 mjeru (46,44%), dok su slučajne šume postigle najveću točnost (84,03%). Arshad Ali [2] koristi podatkovni skup Fortune 1000 iz 2024. godine za predviđanje profitabilnosti poduzeća. Autor provodi eksploratornu analizu podataka te koristi model slučajne šume za klasifikaciju poduzeća kao profitabilnih i neprofitabilnih. Pritom koristi selekciju značajki kako bi prikazao najvažnije značajke prilikom predviđanja. Implementirani model postiže vrlo dobre rezultate s točnosti od čak 99,50% i preciznošću od 99,00%.

Ovaj diplomski rad je strukturiran u četiri poglavlja gdje će se u poglavlju 2. uvesti koncept strojnog učenja te će se opisati modeli korišteni za predviđanje ranga poduzeća. Potom slijedi poglavlje 3., gdje će biti opisan skup podataka Fortune 1000 te provedena eksploratorna analiza. Nakon detaljne analize u poglavlju 4. provest će se implementacija modela stabla odluke, slučajne šume i XGBoosta, zajedno s identifikacijom najvažnih značajki za predviđanje. Za kraj će se u poglavlju 5. vrednovati implementirani modeli i komentirati utjecaj značajki na predviđanje te će se iznijeti osvrt na cjelokupni rad, uz prijedloge za poboljšanja u budućnosti.

2. Općenito o modelima strojnog učenja

Strojno učenje je grana umjetne inteligencije koja se fokusira na omogućavanje računalima i strojevima da oponašaju način na koji ljudi uče, obavljaju zadatke autonomno te poboljšavaju svoje performanse i točnost kroz iskustvo i izloženost većem broju podataka. Zadatak algoritama strojnog učenja je pronaći prirodne uzorke i poveznice u podacima te na temelju toga steći uvid i zatim odlučiti i predviđati. Osnovna podjela strojnog učenja može se svesti na [3]:

- **Nadzirano učenje** (engl. *supervised learning*)
  Algoritmi koriste označene skupove podataka kako bi trenirali podatke s ciljem klasifikacije ili predviđanja vrijednosti. Kod klasifikacije primjeru pridružujemo klasu kojoj taj primjer pripada, a kod regresije neku kontinuiranu vrijednost.

- **Nenadzirano učenje** (engl. *unsupervised learning*)
  Algoritmi analiziraju i grupiraju neoznačene skupove podataka. Tri tipična zadatka nenadziranog učenja su grupiranje podataka (engl. *clustering*), otkrivanje novih vrijednosti ili vrijednosti koje odskaču (engl. *novelty/outlier detection*) i smanjenje dimenzionalnosti (engl. *dimensionality reduction*).

- **Polunadzirano učenje** (engl. *semi-supervised learning*)
  Ovdje se kombinira nadzirano učenje s nenadziranim učenjem. Tijekom treniranja podataka koriste se manji označeni skupovi podataka kako bi se odvijala klasifikacija i ekstrakcija značajki iz većeg, neoznačenog skupa podataka. Koristi se kad ne postoji dovoljno označenih podataka za algoritme nadziranog strojnog učenja.

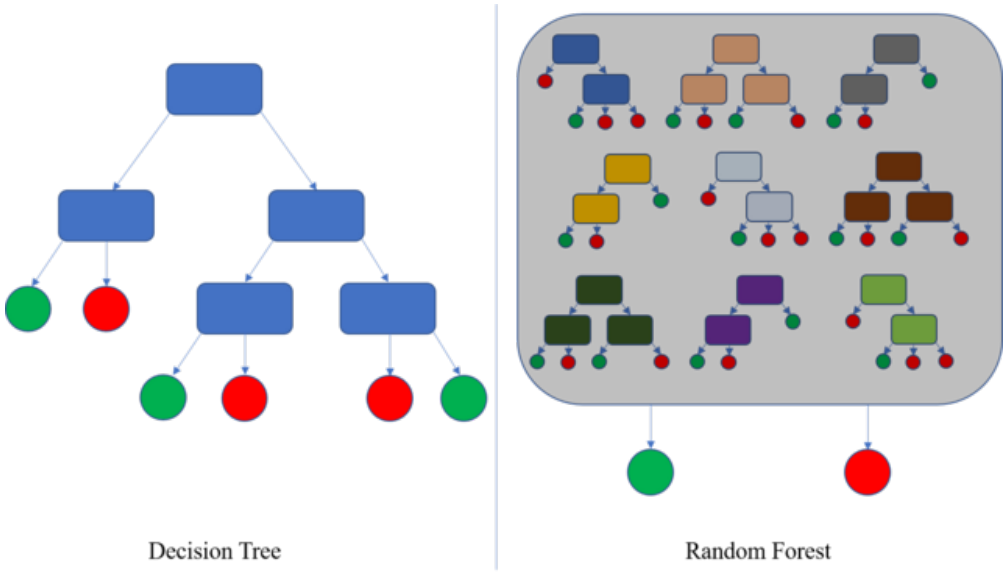
U ovom radu će se za predviđanje uspješnosti poduzeća koristiti sljedeći algoritmi: stablo odluke (engl. *Decision Tree*, DT), slučajna šuma (engl. *Random Forest*, RF) i *Ex-*

treme Gradient Boosting (XGBoost). Osim toga, za imputaciju nedostajućih vrijednosti u podatkovnom skupu prije same implementacije navedenih modela koristit će se i algoritam KNN. Svi navedeni algoritmi pripadaju nadziranom učenju i opisani su u nastavku.

2.1. Stablo odluke i slučajna šuma

Stablo odluke je hijerarhijska stablasta struktura, koja se sastoji od čvorova koji predstavljaju testove značajki, grane predstavljaju izlaze ovih testova, a listovi predstavljaju konačne izlaze odnosno predviđanja (klase ili kontinuirane vrijednosti). [4]

Međutim, pojedinačna stabla odlučivanja mogu biti sklona prenaučenosti (engl. *overfitting*), što može smanjiti njihovu sposobnost generalizacije na nove, nepoznate podatke. Kao rješenje ovog problema razvijen je algoritam slučajne šume, koji sadržava više stabala odluke na različitim podskupovima podataka, pri čemu uzima prosjek kako bi poboljšao točnost predviđanja na odabranom podatkovnom skupu. Osim da se oslanja na odluku samo jednog stabla odluke, RF dakle uzima u obzir predviđanja od više stabala i na temelju predviđanja koje broji "najviše glasova" daje konačnu odluku, što je vizualno prikazano na Slici 2.1. [5]



Slika 2.1. Stablo odluke vs. slučajna šuma [6]

RF je metoda strojnog učenja koja se može koristiti i za klasifikaciju i za regresiju. Zasniva se na konceptu ansambla, što označava proces kombiniranja više klasifikatora

ili regresora kako bi se riješio složen problem i poboljšala izvedba modela. Preciznije, RF se temelji na *baggingu*, što uključuje treniranje više modela na međusobno neovisnim skupovima podataka te donošenjem odluke uzimanjem prosjeka njihovih predviđanja [5].

Algoritam slučajne šume se može opisati u nekoliko koraka [5]:

1. Odaberite nasumičnih K podataka iz skupa za treniranje.
2. Izgradite stabla odluke na temelju odabranih podataka (podskupova).
3. Odaberite broj N za stabla odluke koja želite izgraditi.
4. Ponovite korake 1 i 2.
5. Za nove podatke pronađite predviđanja svakog stabla odluke i dodijelite nove podatke klasi koja dobije većinu glasova u slučaju klasifikacije ili dodijelite novom podatku vrijednost koja je prosjek svih izlaza u slučaju regresije.

2.2. XGBoost

Gradient Boosting Decision Trees (GBDT) je algoritam za učenje skupova stabala odluke sličan slučajnoj šumi, koji se koristi za klasifikaciju i regresiju. I RF i GBDT grade model koji se sastoji od više stabala odluke, no razlika je u tome kako se stabla grade i kombiniraju [7].

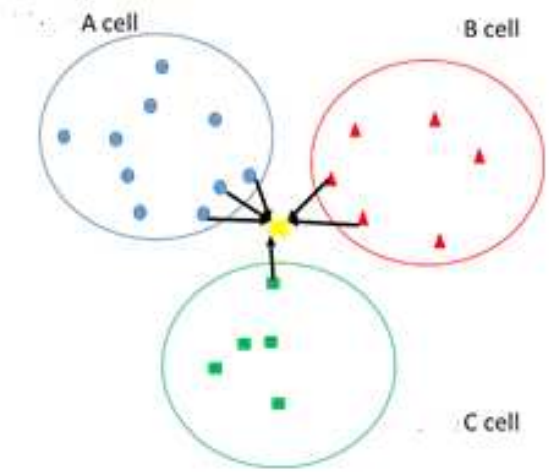
Za razliku od RF-a koje koristi već spomenutu tehniku *bagging*, izraz *gradient boosting* dolazi od ideje o *boostingu* ili poboljšanju jednog slabog modela sekvencijalnim kombiniranjem s drugim slabim modelima. *Gradient boosting* je proširenje *boostinga*, gdje je proces aditivnog generiranja slabih modela ostvaren kao algoritam gradijentnog spusta na ciljnu funkciju. *Gradient boosting* postavlja ciljane izlaze za sljedeći model s ciljem minimizacije pogreške. Ciljani ishodi za svaki slučaj temelje se na gradijentu pogreške (otuda naziv *gradient boosting*) u odnosu na predviđanje [7].

XGBoost je skalabilna i vrlo precizna implementacija *gradient boostinga*, a izgrađena je uglavnom za poboljšanje izvedbe modela strojnog učenja i računalne brzine. S XGBoostom, stabla se grade paralelno, umjesto sekvencijalno kao u GBDT-u. Slijedi strategiju

po razinama, prolazeći kroz vrijednosti gradijenta i koristeći te parcijalne zbrojeve za procjenu kvalitete podjela na svakom mogućem grananju u skupu podataka za treniranje [7].

2.3. K-najbližih susjeda

K-najbližih susjeda je vrlo često korišten algoritam strojnog učenja, koji je poznat po svojoj jednostavnosti i lakoći korištenja. Ne podrazumijeva nikakve pretpostavke o distribucijama podataka, radi i na numeričkim i na kategoričkim podacima. KNN radi na način da uzima K najbližih susjeda neke podatkovne točke na temelju neke od mjera udaljenosti, najčešće euklidske udaljenosti. Klasa ili vrijednost podatkovne točke se zatim određuje na temelju ili prosjeka ili većinskog broja glasova K susjeda, što možemo vidjeti na Slici 2.2. [8] .



Slika 2.2. Primjer algoritma KNN [9]

3. Opis i analiza podatkovnog skupa

3.1. Općenito o Fortune 1000

Podatkovni skup koji će se koristiti za analizu i predviđanje u nastavku rada je Fortune 1000. Skup sadrži detaljne podatke o tisuću poduzeća u Sjedinjenim Američkim Državama s ostvarenim najvećim prihodima. Poduzeća na ovom popisu čine više od dvije trećine gospodarstva zemlje. Podatke o najuspješnijim poduzećima svake godine objavljuje časopis Fortune, renomirani američki poslovni časopis poznat po svom fokusiranju na teme povezane s poslovanjem, ekonomijom, liderstvom, inovacijama i financijama. Časopis je jedan od najprepoznatljivijih izvora informacija u poslovnom svijetu i često se koristi kao referenca za analize i rangiranja poduzeća [10].

Osim popisa Fortune 1000 postoji i prestižniji popis Fortune 500 koji umjesto tisuću rangira petsto najuspješnijih poduzeća. Često se radije koristi Fortune 500 popis jer se smatra da se na Fortune 1000 popisu često mijenjaju poduzeća koje se nalaze pri dnu popisa. S obzirom na to da je Fortune 500 samo podskup većeg popisa Fortune 1000, u radu će se analizirati podaci na temelju svih 1000 poduzeća [10].

3.2. Eksploratorna analiza podataka

Eksploratorna analiza podataka je ključni korak u podatkovnoj znanosti koji prethodi implementaciji modela strojnog učenja. Ona uključuje analizu i vizualizaciju podatkovnog skupa s ciljem boljeg razumijevanja osobitosti podataka, skrivenih uzoraka i potencijalnih odnosa između značajki. U radu će se za provedbu analize i implementaciju modela koristiti programski jezik *Python* s pripadajućim bibliotekama za podatkovnu znanost: *Pandas*, *Numpy*, *Scikit-learn*, *Matplotlib*, *Seaborn* i dr.

3.2.1. Preuzimanje i učitavanje podataka

Podatkovni skup preuzet je u formatu .csv datoteke s javno dostupne platforme za podatkovnu znanost Kaggle[1], a u njemu su poduzeća opisana pomoću 32 značajke, koje predstavljaju financijske, kvalitativne i demografske karakteristike. Značajke su navedene i opisane u Tablici 3.1., gdje su navedeni i pripadajući tipovi značajki. Treba naglasiti da sve značajke koje se odnose na promjenu podrazumijevaju promjenu u odnosu na prethodnu godinu, a novčane vrijednosti su izražene u američkim dolarima ($).

Tablica 3.1. Opisi i tipovi značajki

Značajka	**Opis**	**Tip**
Rank	Rang na ljestvici	Numerička
Company	Naziv poduzeća	Kategorička
Ticker	Dionički simbol	Kategorička
Sector	Sektor	Kategorička
Industry	Industrija	Kategorička
Profitable	Poduzeće je/nije profitabilno	Kategorička
Founder_is_CEO	Osnivač je/nije ujedno i CEO	Kategorička
FemaleCEO	CEO je/nije žena	Kategorička
Growth_in_Jobs	Porast broj zaposlenika	Kategorička
Change_in_Rank	Promjena u rangu	Numerička
Gained_in_Rank	Poduzeću je/nije porastao rang	Kategorička
Dropped_in_Rank	Poduzeću je/nije pao rang	Kategorička
Newcomer_to_the_Fortune_500	Poduzeće je/nije u Fortune 500	Kategorička
Global500	Poduzeće je/nije u top 500 svjetskih poduzeća	Kategorička
Worlds_Most_Admired_Companies	Poduzeće je/nije svrstano u najcjenjenija poduzeća na svijetu	Kategorička
Best_Companies_To_Work_ For	Poduzeće je/nije svrstano u najbolje poslodavce	Kategorička
Number_of_Employees	Broj zaposlenika	Numerička
MarketCap_March28_M	Tržišna kapitalizacija prema 28. ožujku iste godine	Numerička
Revenues_M	Prihod (u milijunima)	Numerička
RevenuePercentChange	Postotna promjena prihoda	Numerička
Profits_M	Dobit (u milijunima)	Numerička
ProfitsPercentChange	Postotna promjena dobiti	Numerička
Assets_M	Imovina (u milijunima)	Numerička
CEO	Chief Executive Officer (izvršni direktor)	Kategorička
Country	Država	Kategorička
HeadquartersCity	Grad sjedišta	Kategorička
HeadquartersState	Savezna država sjedišta	Kategorička
Website	Poveznica na web stranicu	Kategorička
CompanyType	Tip poduzeća - javno/privatno	Kategorička
Footnote	Fusnota	Kategorička
MarketCap_Updated_M	Ažurirana tržišna kapitalizacija (u milijunima)	Numerička
Updated	Datum posljednjeg ažuriranja	Kategorička

[1] `https://www.kaggle.com/datasets/jeannicolasduval/2024-fortune-1000-companies`

Učitana su tri skupa podataka, a svaki skup podataka rangira poduzeća na temelju rezultata za prethodnu fiskalnu godinu. Imamo podatke iz 2024., 2023., i 2022. godine. U radu će se prikazati proces analize podataka primarno za skup iz 2024. godine, ali isto će se primijeniti na podacima za 2023., dok za skup iz 2022. nisu dostupne sve potrebne značajke pa će se koristiti samo kao pomoćni skup za izračun značajki promjena za 2023. godinu.

3.2.2. Odabir značajki i čišćenje podataka

Podatkovni skup sadrži 32 značajke, no neće nam sve one trebati u daljnjoj analizi. Značajke koje možemo izbaciti su sljedeće: *Ticker*, *CEO*, *Country*, *Website*, *Footnote*, *Updated*, *Newcomer_to_the_Fortune500*. *Ticker*, *CEO*, *Website* se odnose na nazivlje koje dalje neće biti potrebno, podatak o državi *Country* je suvišan jer znamo da su sva poduzeća locirana u SAD-u, *Updated* se može zanemariti jer znamo da se radi o podacima za prethodnu fiskalnu godinu, a *Newcomer_to_the_Fortune500* je redundantna značajka jer postoje ostale bitnije značajke povezane s rangom. Nakon izbacivanja navedenih ostaje ukupno 25 značajki.

Sljedeći korak je očistiti podatke. Najprije je potrebno izbaciti duplikate ako oni postoje. U ovom podatkovnom skupu nema duplikata, stoga ne moramo uklanjati nikakve retke.

Zatim je potrebno provjeriti postoje li nedostajuće vrijednosti. Većina modela strojnog učenja ne podržava nedostajuće vrijednosti, zato treba znati kako pravilno postupiti s njima. Nakon provjere nedostajućih vrijednosti, dobivamo da se one nalaze u nekoliko različitih značajki: *MarketCap_March28_M* (41), *RevenuePercentChange* (6), *Profits_M* (2), *ProfitsPercentChange* (150), *MarketCap_Updated_M* (37).

Postoji nekoliko uvriježenih načina kako postupati s nepostojećim vrijednostima, a to su sljedeći [11]:

- zamjena artimetičkom sredinom, medijanom ili modom
- uklanjanje zapisa gdje postoji nedostajuća vrijednost
- zamjena prethodnom/narednom vrijednosti

- interpolacija na temelju susjednih podataka
- zamjena konstantom (većinom *0* ili *-1*)
- predviđanje vrijednosti pomoću metoda strojnog učenja

Što se tiče značajki *MarketCap_March28_M* i *MarketCap_Updated_M*, kreirat ćemo novu značajku koja će objediniti obje odnosno sadržavat će posljednju ažuriranu tržišnu vrijednost za godinu na koju se odnosi. Podatak o tržišnoj kapitalizaciji većinom nedostaje kod poduzeća koja su privatna. Ove retke nećemo samo izbaciti jer tako gubimo podatke o privatnim poduzećima. U ovom slučaju ćemo imputirati nedostajeće vrijednosti pomoću algoritma KNN koji je prethodno opisan u radu. Na taj način vrijednost tržišne kapitalizacije na mjestu nedostajućih vrijednosti postaje jednaka prosjeku *K=5* najbližih susjeda. Ovu metodu koristimo jer je fleksibilna (sami odabiremo vrijednost za *K*), održava veze među podacima i unapređuje točnost [12].

U slučaju nedostajućih vrijednosti kod značajke *Profits_M*, ta dva retka možemo izbaciti jer ih je zanemarivo malen broj.

Kod značajki *RevenuePercentChange* i *ProfitsPercentChange* uzet ćemo u obzir podatke iz prethodne godine. Za to već imamo učitan podatkovni skup Fortune 1000 iz 2023. godine. U već postojeći podatkovni skup dodajemo privremeno dvije značajke u kojima će se nalaziti vrijednosti prihoda (*Revenues_M_2023*) i dobiti (*Profits_M_2023*) za prethodnu godinu. Zatim ćemo na mjestima gdje postotne promjene odnosno *RevenuePercentChange* i *ProfitsPercentChange* nedostaju izračunati te promjene koristeći formule 3.1 i 3.2

$$RevenuePercentChange = \frac{Revenues\_M - Revenues\_M\_2023}{|Revenues\_M\_2023|} \times 100 \tag{3.1}$$

$$ProfitsPercentChange = \frac{Profits\_M - Profits\_M\_2023}{|Profits\_M\_2023|} \times 100 \tag{3.2}$$

Na ovaj način smo uklonili većinu nedostajućih vrijednosti za značajke *RevenuePercentChange* i *ProfitsPercentChange*. Zatim uklanjamo privremene značajke *Revenues_M_2023*

i *Profits_M_2023*, a preostale nedostajeće vrijednosti mijenjamo vrijednosti nula pretpostavljajući da nije bilo promjena u prihodima i dobitima kod tih poduzeća.

Kako su financijske performanse izražene u milijunima, pomnožit ćemo njihove vrijednosti s tisuću kako bi imali apsolutne iznose, koji su većinom u milijardama.

Osim čišćenja podataka i uklanjanja značajki, moguće je dodati nove značajke ako će one biti u budućnosti. U ovom slučaju, osim apsolutnog broja zaposlenika poduzeća *Number_of_Employees* i binarne značajke koja govori je li narastao broj zaposlenika *Growth_in_Jobs*, dodat ćemo i novu značajku *EmplPercentChange*. Nova značajka sadržavat će podatak o postotnoj promjeni zaposlenika poduzeća u odnosu na prethodnu godinu. Značajka će biti bitna jer će prikazivati relativnu promjenu broja zaposlenika uzevši u obzir veličinu poduzeća, a ne samo apsolutnu promjenu. Najprije je potrebno dodati privremenu značajku, koja će sadržavati broj zaposlenika za prethodnu kalendarsku godinu, što dobivamo iz skupa podataka Fortune 1000 za 2022. godinu. Zatim vrijednosti nove značajke dobivamo pomoću formule 3.3

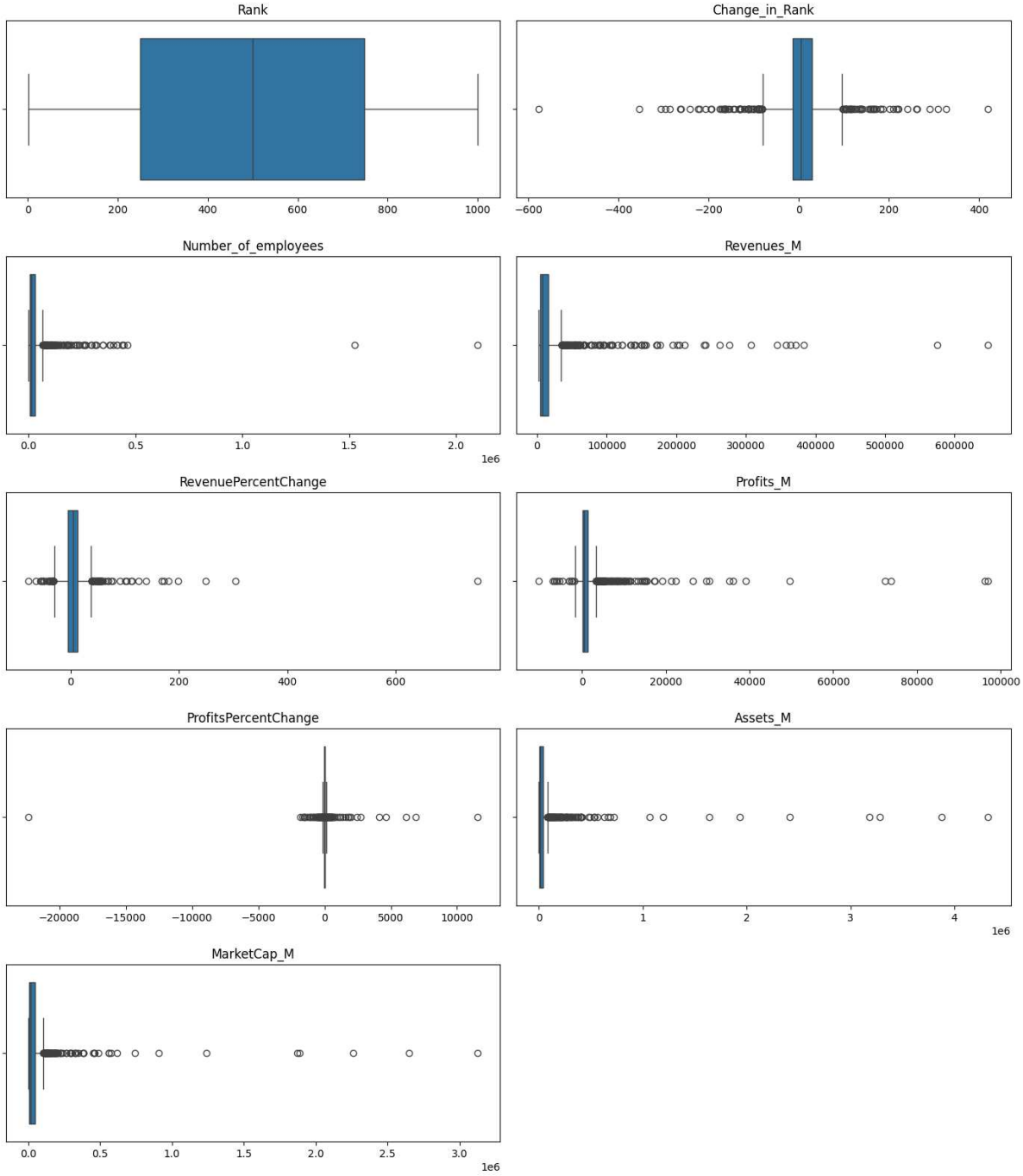
$$EmplPercentChange = \frac{Number\_of\_Employees - Prev\_Num\_of\_Employees}{Prev\_Num\_of\_Employees} \times 100 \qquad (3.3)$$

Nakon stvaranja nove značajke više nam ne treba privremena značajka za prethodni broj zaposlenika *Prev_Num_of_Employees* pa ju uklanjamo. Zatim provjeravamo novonastale nedostajuće vrijednosti u značajki *EmplPercentChange*. Dobivamo da postoji 77 nedostajućih vrijednosti, što ćmo riješiti zamjenom s vrijednosti *0*, pretpostavljajući da nema promjene broja zaposlenih za ta poduzeća.

Stvorit ćemo još jednu novu značajku *Rank_Change* koja će poprimiti vrijednosti ovisno o značajkama *Gained_in_Rank* i *Dropped_in_Rank*. Poduzeća će za ovu značajku poprimiti upravo vrijednosti *Gained_in_Rank*, *Dropped_in_Rank* i *Same_Rank*. Osim toga, dodat ćemo i značajku *Year* koja će označavati iz koje je godine skup podataka.

Nakon uklanjanja nedostajućih podataka i dodavanja/uklanjanja značajki, provjeravamo postoje li stršeće vrijednost u podacima. Stršeće vrijednosti prikazujemo pomoću kutijastih dijagrama (engl. *boxplot*) za numeričke značajke. Vizualizacije se mogu vidjeti

na Slici 3.1.



Slika 3.1. Provjera stršećih vrijednosti pomoću kutijastih dijagrama

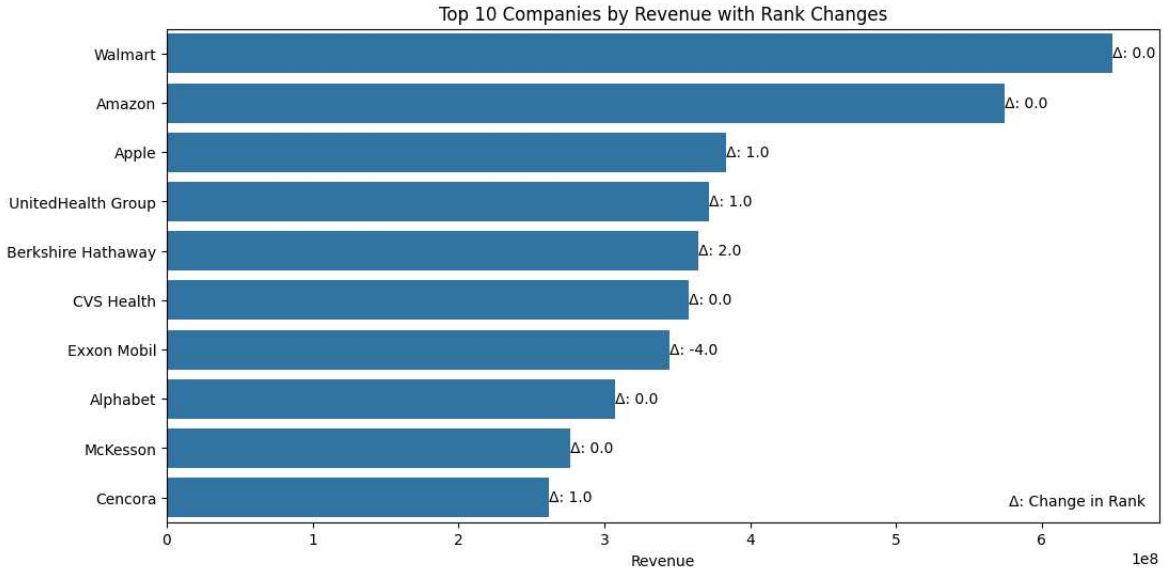
Iz prikazanih dijagrama možemo uočiti da sve numeričke značajke osim ranga sadrže stršeće vrijednosti. Iako su ove vrijednosti prisutne, odlučili smo ih ne uklanjati, jer se radi o financijskim podacima, gdje su poduzeća koja odstupaju od prosjeka često od velike važnosti za analizu.

3.2.3. Vizualizacije podataka i priprema za modele

U ovom poglavlju prikazat će se zanimljive vizualizacije podataka iz kojih će se potencijalno doći do određenih zaključaka korisnih za daljnju analizu. Vizualizacije se odnose za skup iz 2024. godine, osim ako nije drugačije naznačeno.

Numeričke značajke

Najprije pomoću stupčastog dijagrama (engl. *bar chart*) prikazujemo prvih deset poduzeća i promjene njihovih rangova u odnosu na prethodnu godinu, što se može vidjeti na Slici 3.2.



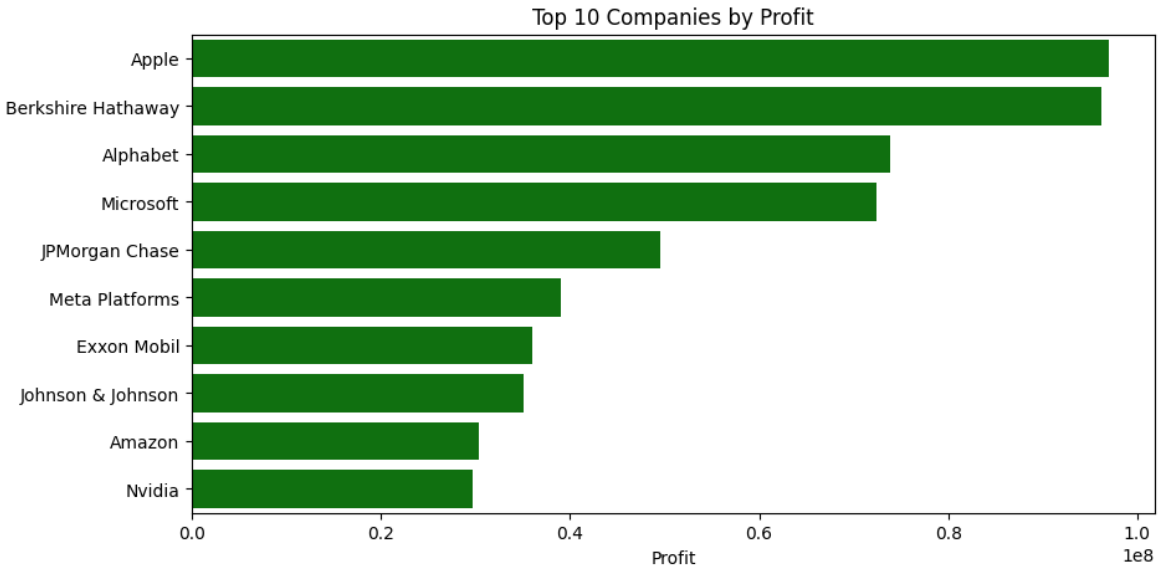
Slika 3.2. Prvih deset poduzeća prema ostvarenim prihodima i njihovi rangovi

Vidimo da se poredak pri vrhu ljestvice nije znatno mijenjao. Iz dijagrama iščitavamo da je prvih deset poduzeća prema najvećim ostvarenim prihodima redom:

1. Walmart
2. Amazon
3. Apple
4. UnitedHealth Group
5. Berkshire Hathaway
6. CVS Health
7. Exxon Mobil

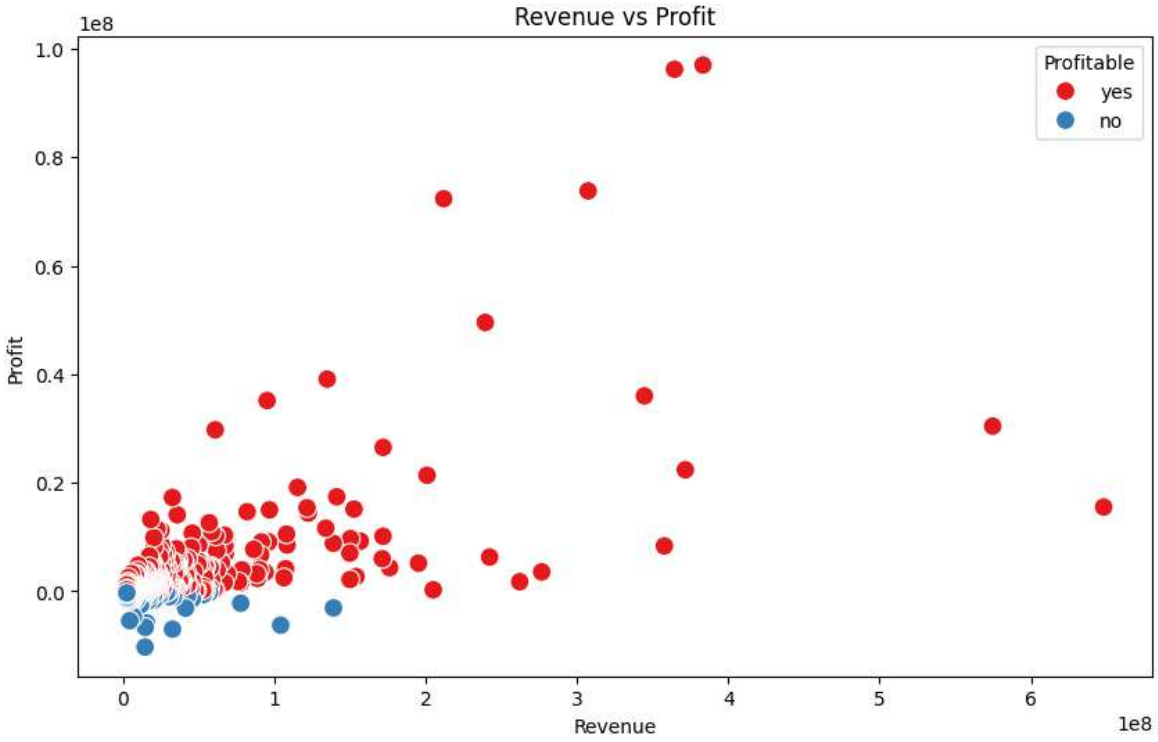
8. Alphabet
9. McKesson
10. Cencora

Za usporedbu ćemo prikazati i prvih deset poduzeća prema najvećoj ostvarenoj dobiti. Treba napomenuti da je razlika između te dvije značajke što se prihod odnosi se na ukupnu količinu novca koju poduzeće donosi prodajom, dok je dobit iznos novca koji poduzeće zaradi nakon odbitka svih troškova. Stupčasti dijagram prikazan je na Slici 3.3.



Slika 3.3. Prvih deset poduzeća prema ostvarenoj dobiti

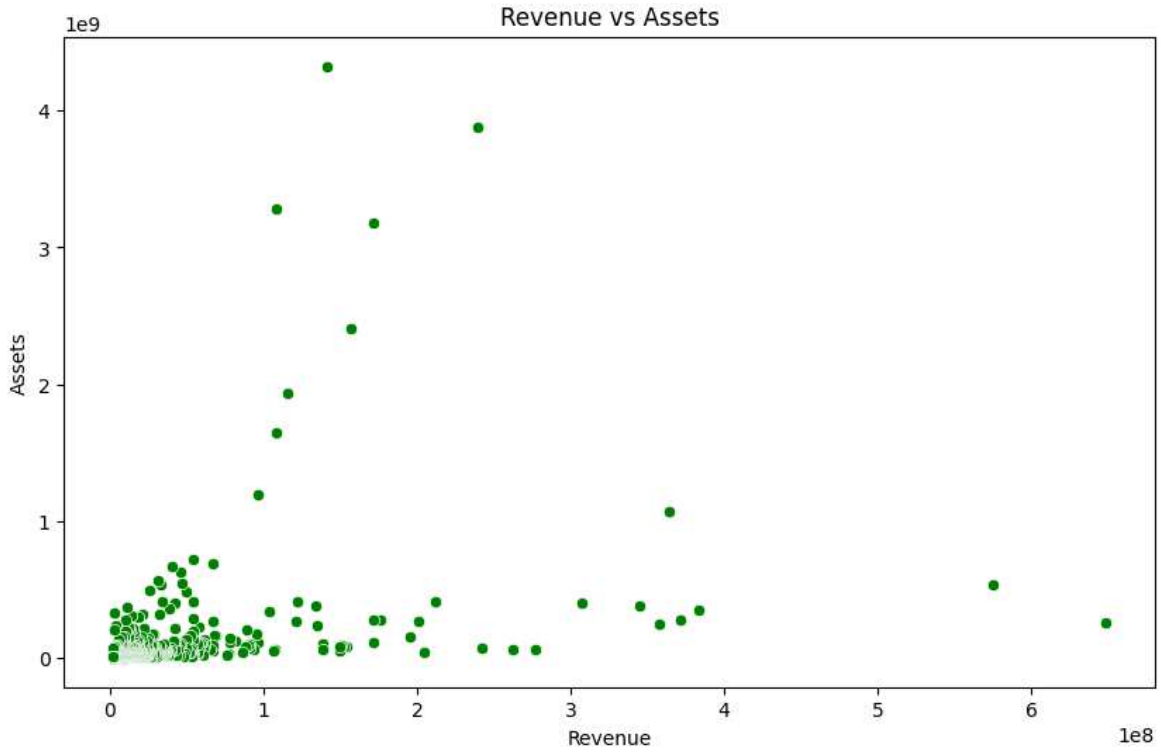
Iz dijagrama vidimo da poredak nije jednak kao na prikazu najvećih ostvarenih prihoda, zbog čega ćemo prikazati povezanost između dobiti i prihoda pomoću dijagrama raspršenosti (engl. *scatter plot*). Na Slici 3.4. prikazan je odnos između prihoda i dobiti, gdje dodatno boja označava je li poduzeće profitabilno.



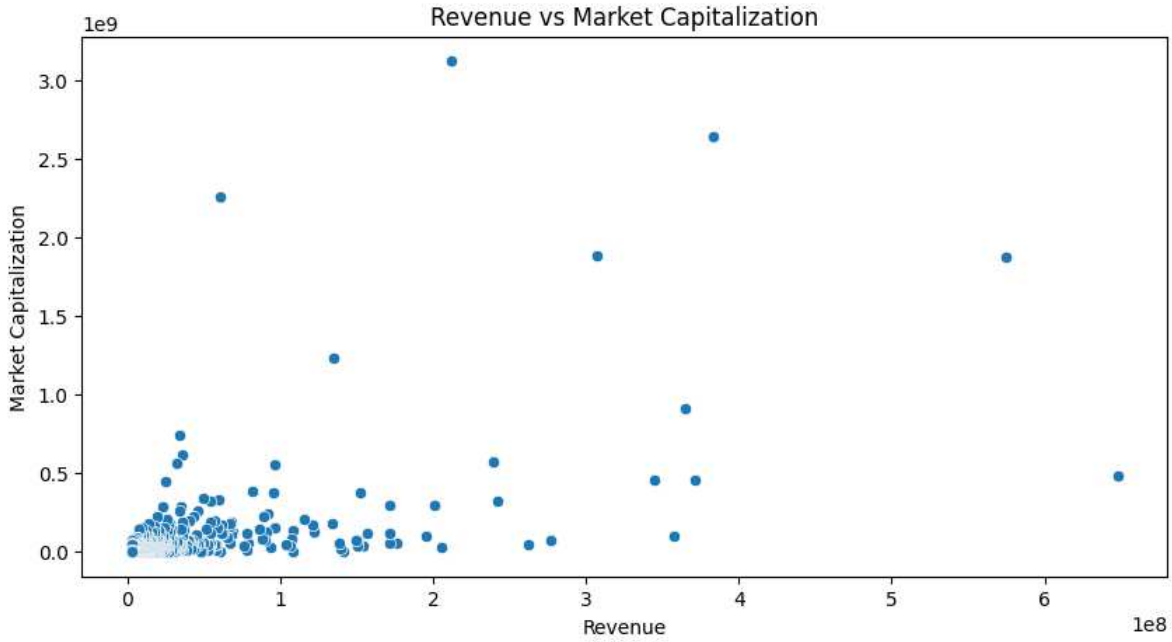
Slika 3.4. Dijagram raspršenosti za prihode i dobiti

Iz dijagrama se vidi ovisnost prihoda i dobiti, no ona nije linearna. Neka poduzeća, iako imaju visoke prihode, nisu profitabilna.

Nakon toga zanima nas kako je povezan prihod s imovinom poduzeća i tržišnom kapitalizacijom. To također prikazujemo pomoću dijagrama raspršenosti na Slici 3.5. i Slici 3.6.



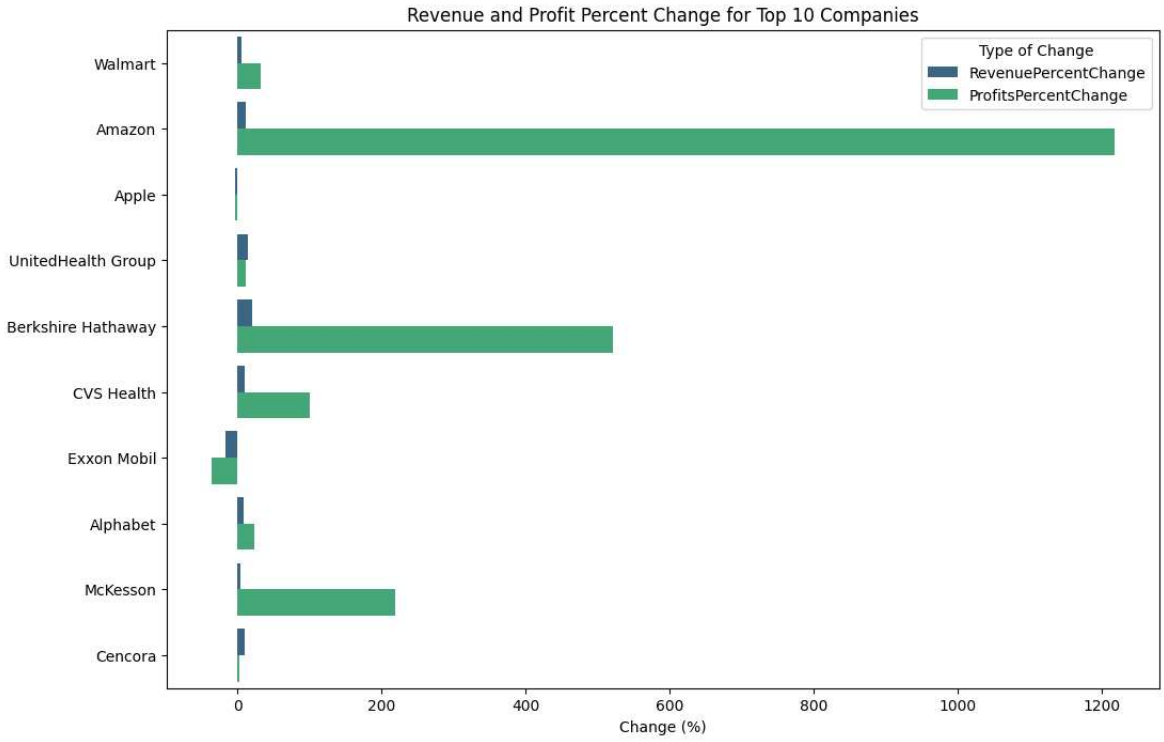
Slika 3.5. Dijagram raspršenosti za prihode i imovinu



Slika 3.6. Dijagram raspršenosti za prihode i tržišnu kapitalizaciju

Nakon prikazivanja i uspoređivanja apsolutnih financijskih vrijednosti zanima nas i odnos promjena tih vrijednosti. Prikazat ćemo poduzeća relativne promjene prihoda i dobiti u odnosu na prethodnu godinu za prvih deset poduzeća na Fortune 1000 popisu.

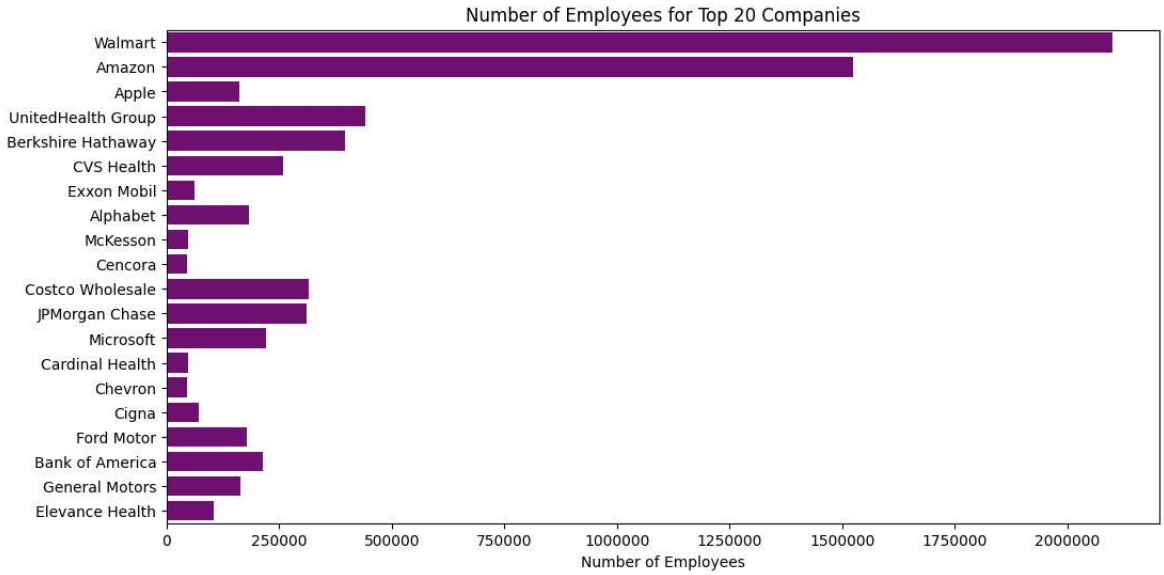
Promjene su prikazane usporednim stupčastim dijagramima na Slici 3.7.



Slika 3.7. Usporedba promjena prihoda i dobiti (u %) za prvih deset poduzeća

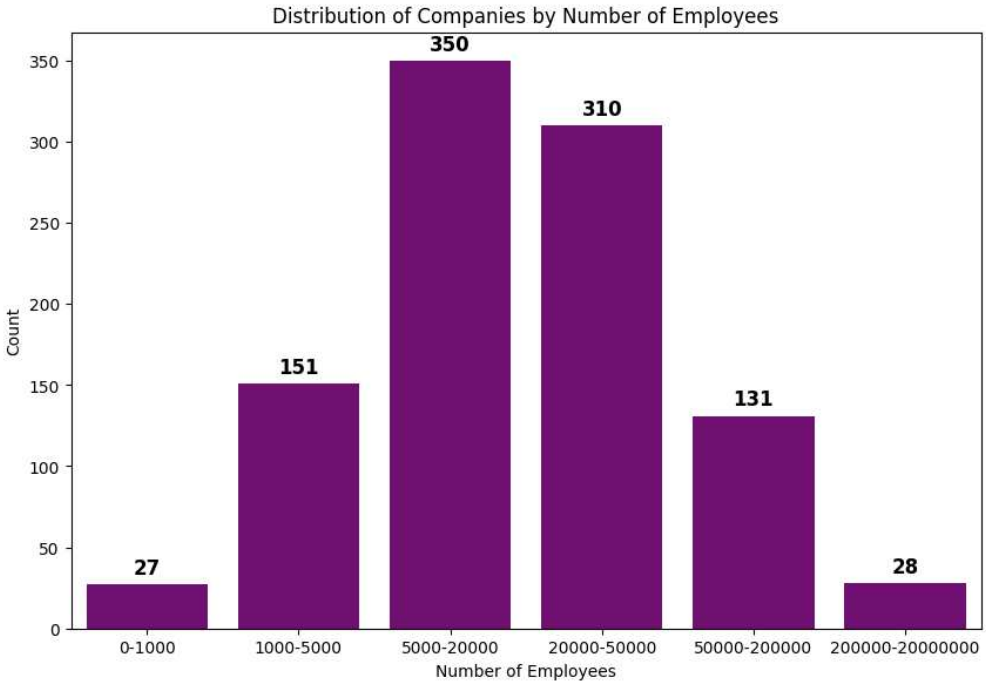
Iz grafa se vidi da promjena prihoda većinom nije znatna za najuspješnija poduzeća, no zato su velike varijacije u promjeni dobiti. Npr. Amazonu je dobit u odnosu na prethodnu fiskalnu godinu narasla više od izuzetnih 1200%, dok se prihod nije uvelike mijenjao. 5.

Nadalje, zanima nas kako je distribuiran broj zaposlenika po poduzećima. Najprije ćemo prikazati broj zaposlenih za najuspješnijih 20 poduzeća na Slici 3.8.



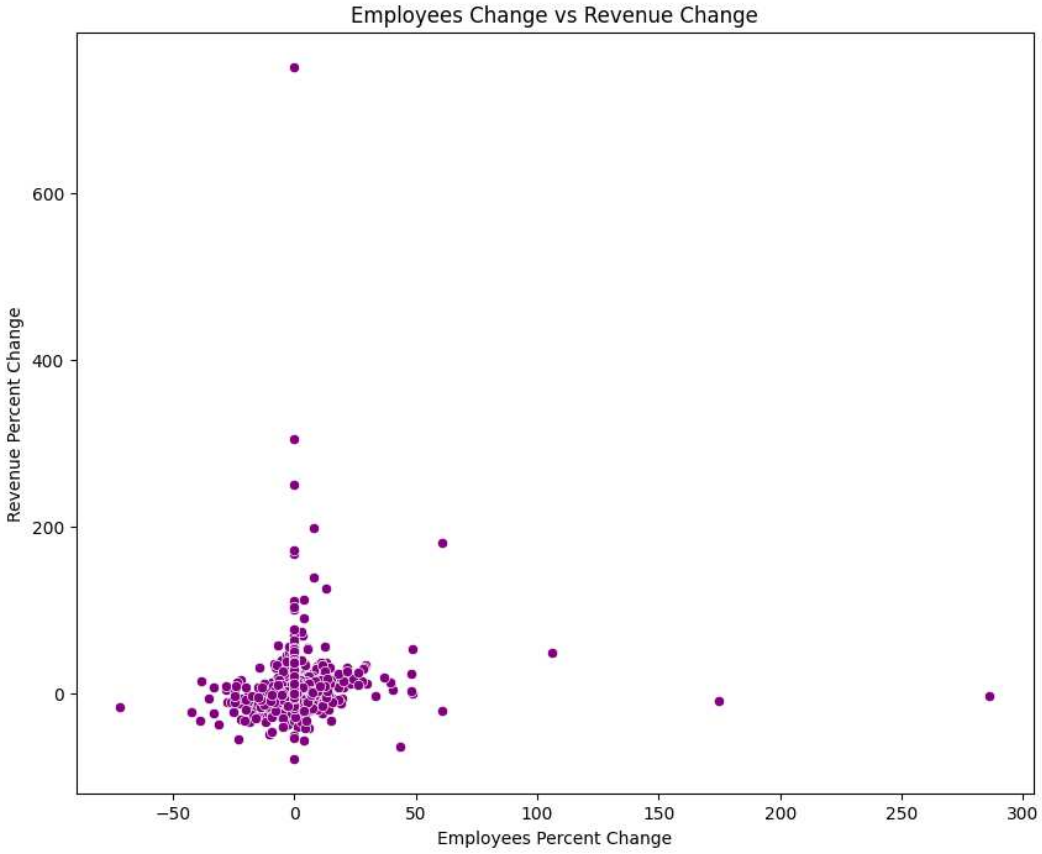
Slika 3.8. Broj zaposlenih za prvih 20 poduzeća

Walmart i Amazon su stršeće vrijednosti uzevši u obzir broj zaposlenih, no iz ostatka vizualnog prikaza se vidi da nije pravilo da poduzeća s najviše zaposlenih ostvaruju najveće prihode. Navedeno se najbolje vidi na primjeru poduzeća Apple. Kako bi bolje razumjeli distribuciju broja zaposlenih po poduzećima, podijelit ćemo poduzeća u šest skupina prema broju zaposlenih: 0-1000, 1000-5000, 5000-20000, 20000-50000, 50000-200000 i 200000-2000000. Histogram je prikazan na Slici 3.8.



Slika 3.9. Distribucija poduzeća prema broju zaposlenih

Provjerit ćemo odnos između rasta broja zaposlenih i rasta prihoda dijagramom raspršenosti na Slici 3.10.



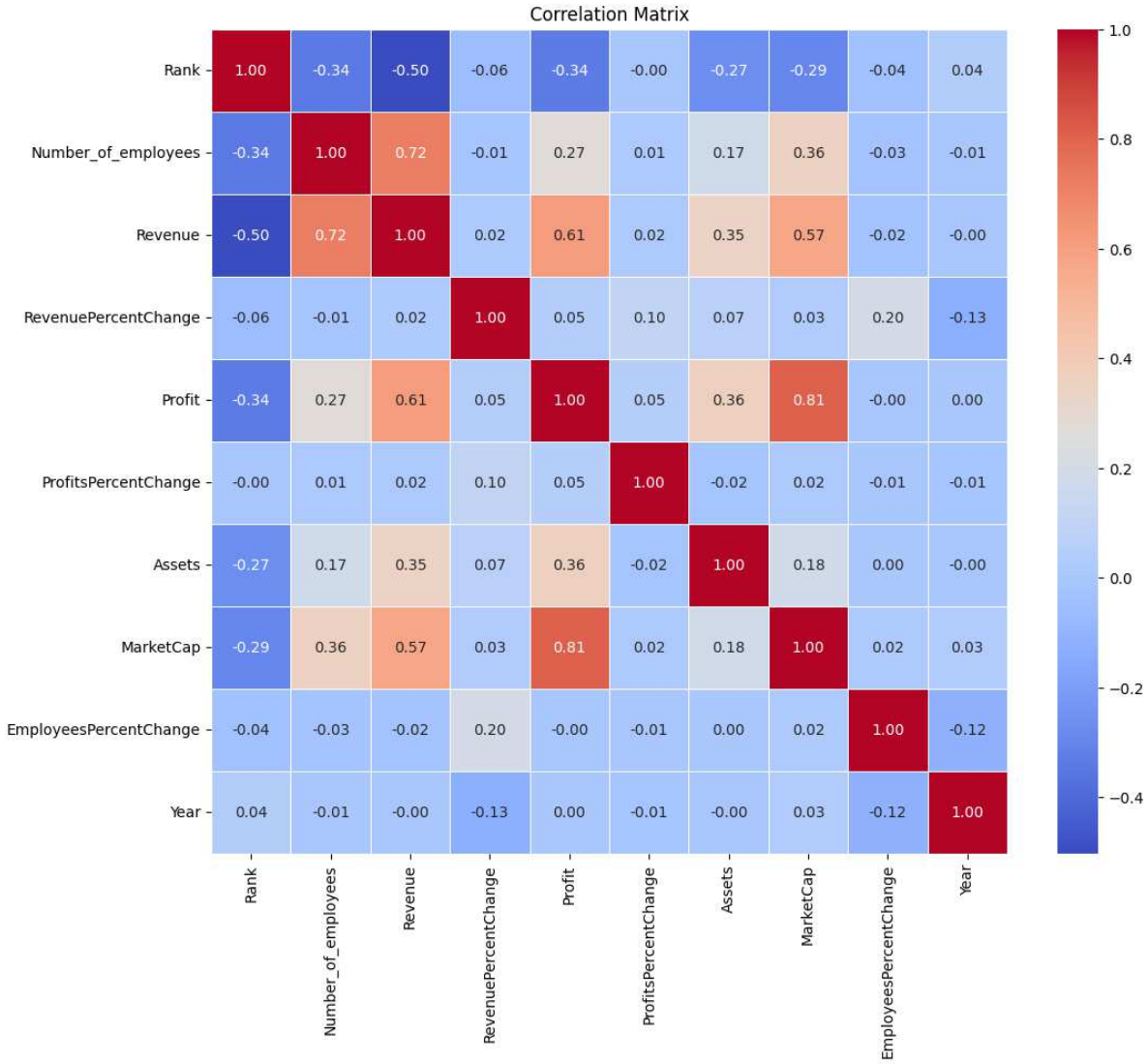
Slika 3.10. Distribucija poduzeća prema broju zaposlenih

Iz vizualnog prikaza vidimo da nema jasne ovisnosti između navedenih dviju značajki, što bi značilo da zapošljavanje novih ljudi ne uvjetuje porast prihoda.

Osim pojedinačnih prikaza značajki i odnosa između dviju značajki, zanima nas i kako su sve značajke međusobno povezane. Za prikaz međuovisnosti svih numeričkih značajki koristit će se korelacijska matrica, gdje su značajke rapoređene u stupce i retke, a koeficijent korelacije poprima vrijednosti između *-1* i *1*, gdje *-1* označava savršenu negativnu korelaciju, *+1* savršenu korelaciju, a *0* znači da nema korelacije između dviju značajki. Za prikaze korelacija koristi se Pearsonov koeficijent koji računa linearnu povezanost između značajki koristeći formulu 3.4

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.4}$$

gdje je n broj podatkovnih točki, x_i i y_i su pojedinačne vrijednosti za promjenjive značajke X i Y, a $\bar{x}$ i $\bar{y}$ srednje vrijednosti promjenjivih značajki X i Y [13]. Korelacijska matrica prikazana je na Slici 3.11.



Slika 3.11. Korelacijska matrica

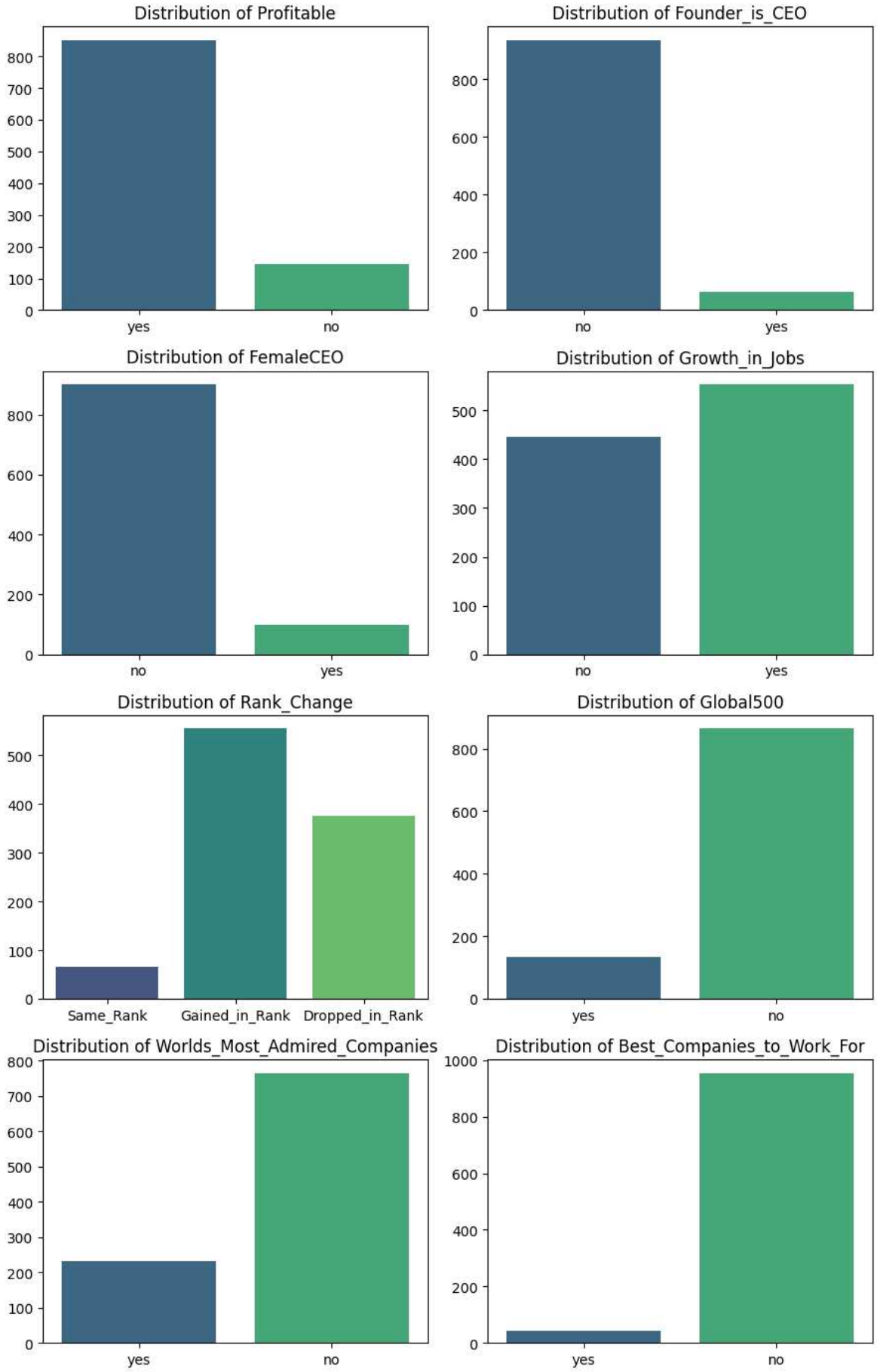
Iz korelacijske matrice možemo vidjeti veliku lineanu povezanost ($|r| > 0.6$) između sljedećih značajki:

- *Number_of_Employees* i *Revenue*
- *Revenue* i *Profit*
- *Profit* i *MarketCap*.

Potrebno je ukloniti značajke koje imaju visok korelacijski koeficijent s drugom značajkom kako ne bi došlo do multikolinearnosti prilikom implementacije modela. Kako znamo da su poduzeća na Fortune 1000 popisu rangirana prema ostvarenim prihodima, uklonit ćemo značajku *Revenue* kako bismo vidjeli koje ostale značajke utječu na rang. Osim toga, uklonit ćemo značajku *MarketCap* zbog visoke povezanosti sa značajkom *Profit*.

Kategoričke značajke

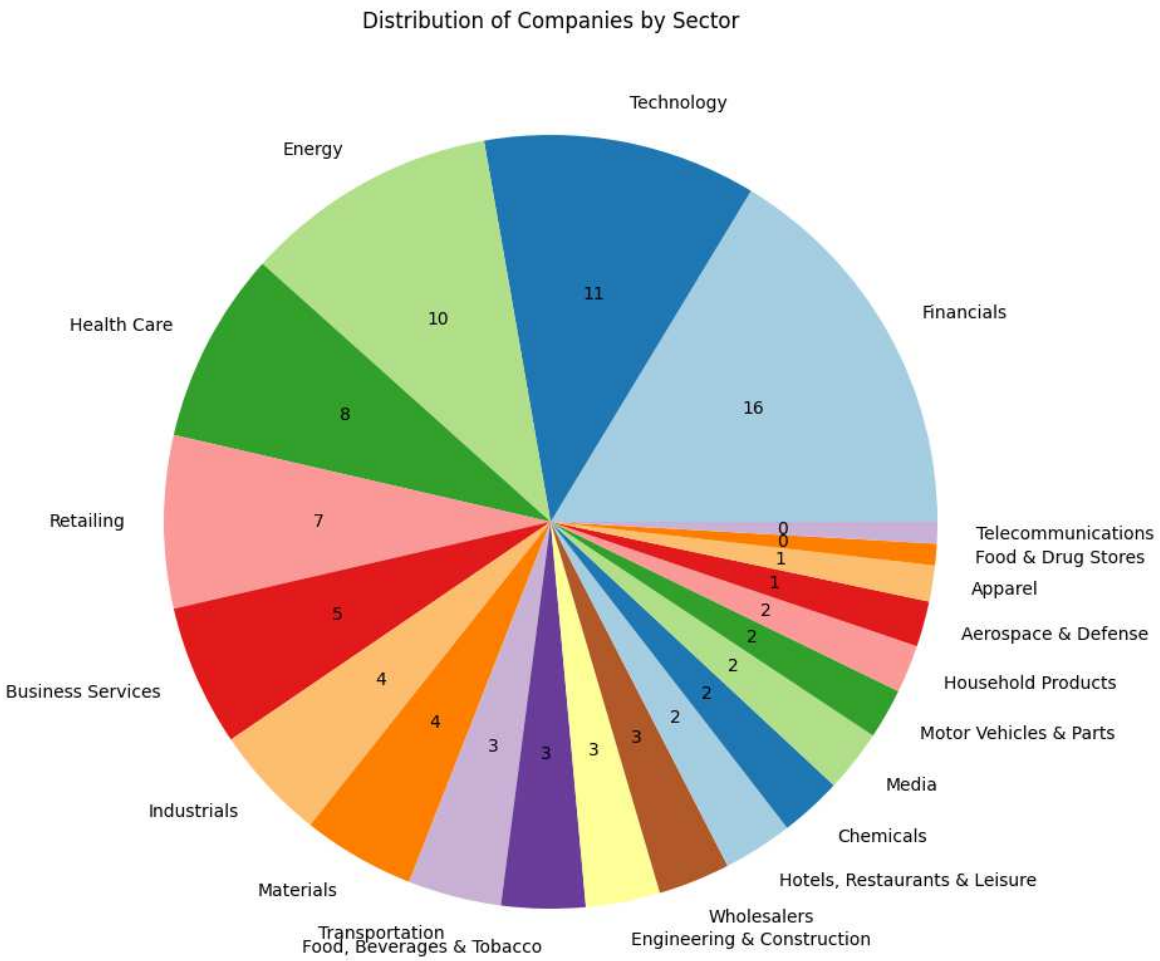
Nakon što smo prikazali numeričke značajke, vizualizirat ćemo i proučiti kategoričke značajke. Prikazat ćemo na nekoliko stupčastih dijagrama distribucije kategorija po značajkama, što je prikazano na Slici 3.12.



Slika 3.12. Distribucije kategoričkih značajki

Iz prikaza se može doći do nekoliko zaključaka: većina poduzeća je profitabilno, većinom osnivači nisu ujedno i izvršni direktori, pretežito su muškarci izvršni direktori. Što se tiče pristutnosti Fortune poduzeća na ostalim prestižnim popisima, manjina njih se nalazi na popisima globalnih 500, najcjenjenija svjetska poduzeća te najbolji poslodavci. Promjena ranga je po poduzećima različita, ali najviše je poduzeća doživjelo porast u rangu u odnosu na prethodnu godinu. Zapošljavanje novih radnika je po poduzećima podjednako prisutno, iako je nešto više poduzeća gdje je narastao broj zaposlenika.

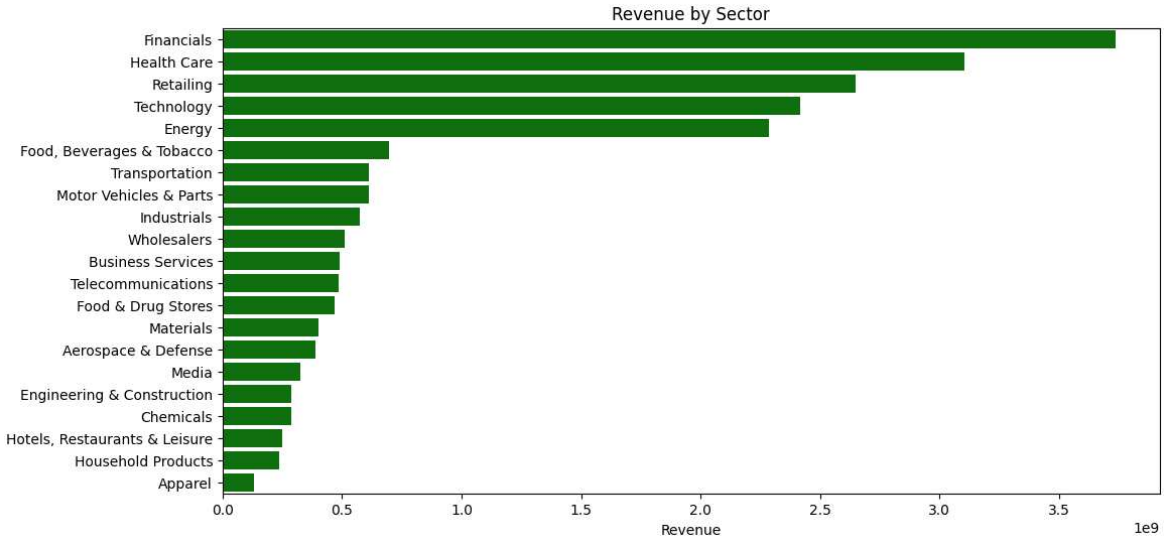
Sljedeće što ćemo prikazati je distribucija poduzeća po različitim sektorima, što je vidljivo na Slici 3.13.



Slika 3.13. Distribucija poduzeća po sektorima

Iz priloženog se vidi da su poduzeća otprilike jednako raspodijeljena po sektorima, s nešto više poduzeća u financijskom, tehnološkom i energetskom sektoru. Zanima nas i

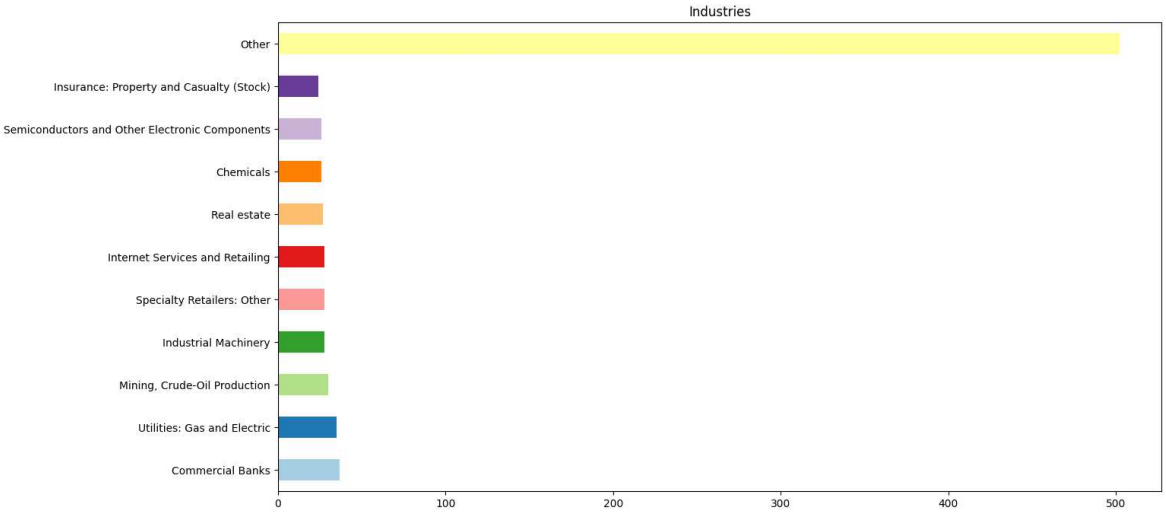
koji sektori ostvaruju najveće prihode, a time su i pripadajuća poduzeća bolje rangirana. Stupčasti dijagram je prikazan na Slici 3.14.



Slika 3.14. Distribucija poduzeća po sektorima

Iz vizualnog prikaza može se doći do zaključka da najviše prihoda ostvaruju financijski, zdravstveni, prodajni, tehnološki i energetski sektor.

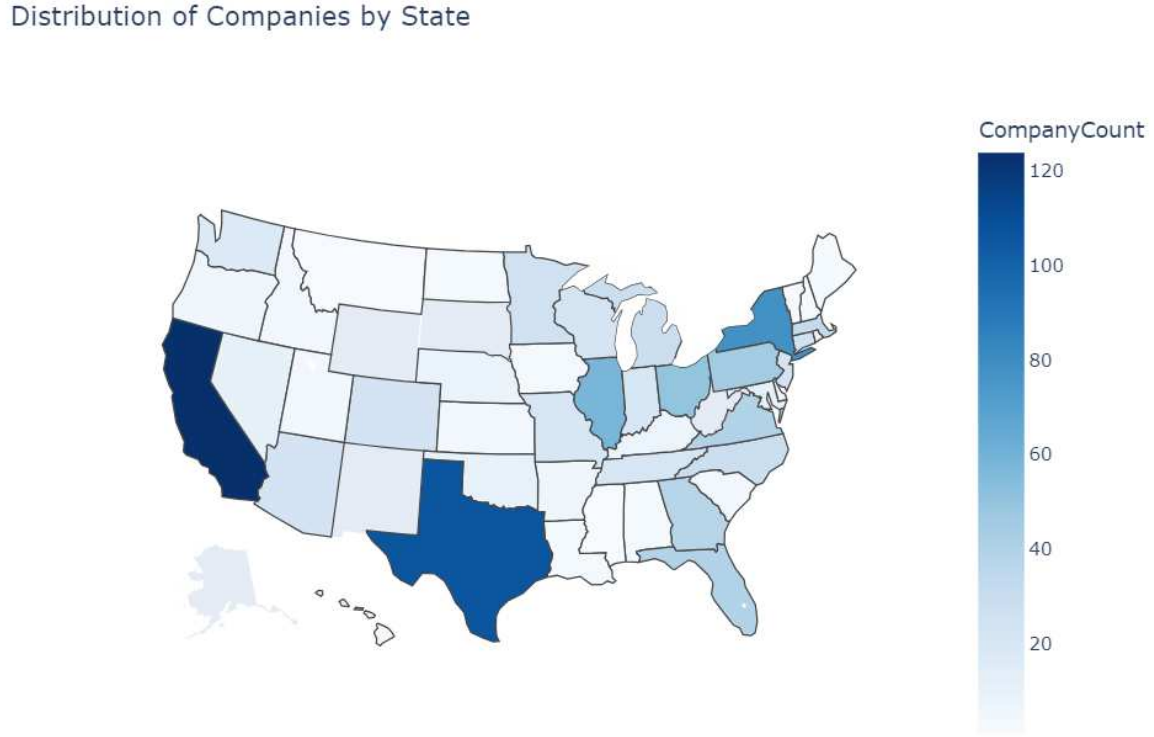
Provjerit ćemo i kakva je distribucija poduzeća po industrijama. Ispisom jedinstvenih vrijednosti pomoću metode *unique()*, dolazimo do zaključka da je industrija uži pojam od sektora te da podatkovni skup sadrži poduzeća iz čak 75 različitih industrija. Provjeravamo ima li prevladavajućih industrija stupčastim dijagramom na Slici 3.15.



Slika 3.15. Distribucija poduzeća po industrijama

Vidimo da nijedna industrija ne broji znatno veći broj poduzeća od ostalih, stoga se može reći da su poduzeća jednoliko distribuirana po industrijama.

Podatkovni skup sadrži i geografske podatke odnosno za svako poduzeće je navedeno u kojoj saveznoj državi i gradu se nalazi njegovo sjedište. Prikazat ćemo prostornu raspodjelu poduzeća na karti SAD-a na Slici 3.16.



Slika 3.16. Distribucija poduzeća po saveznim državama

Iz prikaza na karti vidimo da se najviše poduzeća nalazi u Kaliforniji, Texasu i New Yorku.

Posljednji korak prije implementacije modela strojnog učenja je kodiranje kategoričkih značajki. Naime, većina modela strojnog učenja ne podržava kategoričke vrijednosti, stoga ih je potrebno pretvoriti u numeričke vrijednosti. Za kodiranje će se koristiti *label encoding*, gdje se svakoj kategoriji dodjeljuje jedinstvena numerička vrijednosti, bez postojanja hijerarhije između vrijednosti.

4. Primjena modela strojnog učenja

Glavni cilj ovog rada je implementirati modele za predviđanje uspješnosti poduzeća. Kako je Fortune 1000 jedan od najprestižnijih popisa u svijetu poduzetništva, uspješnost se u ovom slučaju mjeri rangom na navedenoj ljestvici. Poduzeća su rangirana prema ostvarenim prihodima, stoga ćemo značajku *Revenue* izbaciti iz ulaznih značajki. Svrha modela je pokušati što točnije predvidjeti rang poduzeća na temelju ostalih ulaznih značajki te potom utvrditi koje druge značajke implicitno utječu na rang.

Za predviđanje ranga poduzeća na ljestvici Fortune 1000 koristit će se tri različita algoritma: stablo odlučivanja, slučajna šuma i XGBoost. Navedeni modeli su odabrani zato što su relativno robusni na stršeće vrijednosti (kojih je u ovom podatkovnom skupu puno), implicitno podnose kolinearnost značajki, nije potrebna prethodna standardizacija podataka i poznati su po tome što postižu visoke točnosti [14].

Dakle, za svaki od triju modela treba odraditi sljedeće: prvo će se model implementirati sa zadanim hiperparametrima, zatim će se pomoću dvije metode računanja važnosti značajki identificirati najvažnije značajke za predviđanja modela, nakon toga će se provesti optimizacija hiperparametara na validacijskom skupu s reduciranim ulaznim značajkama te će se ponovno evaluirati model na testnom skupu.

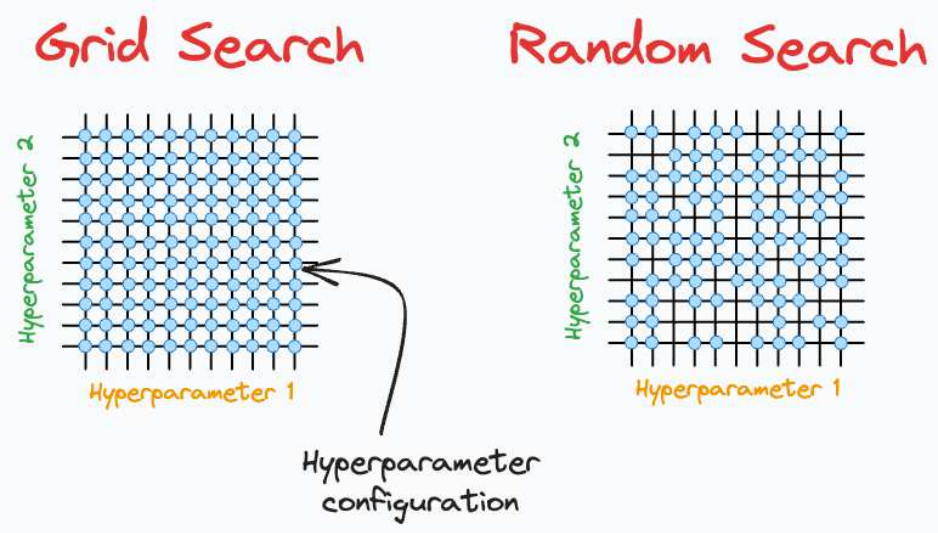
4.1. Implementacija modela predviđanja uspješnosti poduzeća

Završni korak pripreme podataka za implementaciju modela uključuje spajanje skupa podataka iz 2023. i 2024. godine. Dodatno treba izbaciti značajku Worlds_Most_Admired_Companies, koja ne postoji u skupu iz 2023. godine, a ionako nije korisna kao ulazna

značajka u model jer samo daje informaciju o rangu na drugoj ljestvici. Osim toga, izbacit ćemo i značajku *Company* jer se radi o nominalnoj kategoričkoj značajki. Kod sva tri modela će se kao skup za treniranje koristiti skup podataka za 2023. godinu, gdje će se prema povijesnim podacima nastojati vidjeti koje su značajke najviše utjecale na rang te će se to onda koristiti da se predvidi rang na testnom skupu za 2024. godinu. Ulaz u modele su sve značajke osim ranga, dok je ciljna značajka upravo rang. Preostale značajke nakon eksploratorne analize i prilagođavanja naziva su: *Sector, Industry, Profitable, FounderCEO, FemaleCEO, GrowthInJobs, BestCompaniesToWorkFor, NumberOfEmployees, RevenuePercentChange, Profit, ProfitsPercentChange, Assets, HeadquartersCity, HeadquartersState, CompanyType, EmployeesPercentChange, Year*.

Prilikom implementacije modela potrebno je podijeliti podatke na tri dijela: skup za treniranje, validacijski skup i testni skup. Na skupu za treniranje se model uči odnosno na njemu model prepoznaje skrivene uzorke među podacima, na validacijskom skupu se podešavaju hiperparametri modela, a na testnom skupu se vrši konačna evaluacija modela nakon što je on treniran i optimiziran.

Za optimizaciju hiperparametara koristit će se *Random Search*, koji koristi tehniku nasumičnog pretraživanja različitih kombinacija hiperparametara prema zadanom rasponu ili distribuciji. Za razliku od tehnike *Grid Search*, koji hiperparametre pretražuje po *rešetci* i tako isprobava sve moguće kombinacije parametara, *Random Search* je računalno povoljniji jer pronalazi kombinacije nasumično te ih potom evaluira. Razlika između dvaju pristupa prikazana je na Slici 4.1. [15].



Slika 4.1. Grid Search vs Random Search [15]

4.1.1. Regresor stabla odluke i regresor slučajne šume

Uzevši u obzir da je slučajna šuma ansambl više stabala odluke, implementacije ovih dvaju modela su vrlo slične, stoga će se opisati u istom potpoglavlju. Kao što je već spomenuto, za treniranje su se koristili podaci iz 2023. godine, a za testiranje podaci iz 2024. godine. Ulaz u model su sve značajke osim ranga, dok je ciljna značajka upravo rang. Modeli su naprije implementirani sa zadanim vrijednostima u *scikit-learn* biblioteci. Nakon toga je s ciljem poboljšanja izvedbe modela provedena optimizacija hiperparametara, gdje se pomoću tehnike *Random Search* pronalazila optimalna kombinacija hiperparametara. Kako ne bi došlo do prenaučenosti na validacijskom skupu, podešavali su se samo neki hiperparametri. Opisi hiperparametara, njihove zadane vrijednosti i vrijednosti za optimizaciju prikazane su u Tablici 4.1. [16] [17].

Tablica 4.1. Hiperparametri

Hiperparametar	**Opis**	**Zadana vrijednost**	**Vrijednosti za optimizaciju**
max_depth	Maksimalna dopuštena dubina stabla	`None`	`range(2, 51)`
min_samples_split	Najmanji broj uzoraka potreban za podjelu unutarnjeg čvora	`2`	`range(2, 11)`
min_samples_leaf	Najmanji broj uzoraka potreban za čvor lista	`1`	`range(1, 11)`
max_features	Funkcija najvećeg broja značajki koje se uzimaju u obzir prilikom traženja najbolje podjele čvora	`None`	`['sqrt', 'log2', None]`

Kod slučajne je šume uz navedene hiperparametre važan i hiperparametar broj stabala (*n_estimators*). On je postavljen na 100 i neće se optimizirati jer je dovoljno postaviti vrijednosti na velik broj. Osim definiranja prostora hiperparametara, prilikom optimizacije je moguće zadati i druge parametre. Jedan od najvažnijih parametara je broj slojeva za unakrsnu validaciju. Unakrsna validacija (engl. *cross validation*) je tehnika za podjelu podataka, pri čemu se model trenira i potom evaluira k puta na različitim uzorcima. U ovom slučaju je odabran $k=10$, što znači da se model trenira deset puta, gdje svaki sloj jednom služi kao validacijski skup, dok preostali dijelovi čine skup za treniranje. Na kraju se rezultati iz svakog koraka validacije prosječno izračunavaju kako bi se dobila pouzdanija procjena performansi modela [18].

Uz broj slojeva moguće je defnirati i broj kombinacija koje će se isprobati te metrika za procjenu izvedbe unakrsno validiranog modela na testnom skupu. Broj kombinacija je postavljen na 100, a odabrana metrika za procjenu izvedbe je koeficijent determinacije (R^2), što će biti detaljnije pojašnjeno u poglavlju 5.

4.1.2. XGBoost

Kao i kod prethodna dva modela, XGBoost je prvotno pokrenut sa zadanim hiperparametrima, nakon čega se radila optimizacija hiperparametara pomoću tehnike *Random Search*. Opis hiperparametara i njihove vrijednosti prije i tijekom optimizacije se nalaze u Tablici 4.2.

Tablica 4.2. Hiperparametri XGBoosta

Hiperparametar	**Opis**	**Zadana vrijednost**	**Vrijednosti za optimizaciju**
max_depth	Maksimalna dopuštena dubina stabla	`6`	`range(1,11)`
learning_rate	Stopa učenja	`0.3`	`stats.uniform(0.01, 0.1)`
subsample	Poduzorak podataka za svako stablo	`1`	`stats.uniform(0.5, 0.5)`
n_estimators	Broj stabala	`100`	`range(50,201)`

Broj kombinacija je kao i kod dvaju regresora postavljen na 100, a odabrana metrika za procjenu izvedbe je koeficijent determinacije.

4.2. Identifikacija najvažnijih značajki za predviđanje

Nakon implementacije triju modela, cilj nam je identificirati koje su značajke najviše doprinijele predviđanjima modela te onda koristiti samo te značajke za ulaz u model. Za to će se koristiti dvije tehnike: važnost značajki temeljena na nečistoći (engl. *Impurity-based Feature Importance*, IFI) ugrađena u modele stabala u *scikit-learn* biblioteci te permutacijska važnost (engl. *Permutation Importance*, PFI) iz iste biblioteke. Glavna razlika između dvaju pristupa je što se IFI računa na skupu za treniranje, a PFI na testnom skupu. Također, važnosti izračunate metodom IFI zbrajaju se u *1*, dok kod PFI to nije slučaj.

Važnost značajki temeljena na nečistoći koristi vlastite izračune modela za mjerenje važnosti značajki, poput *Gini* važnosti i prosječnog smanjenja točnosti. Važnost pojedine značajke se računa kao ukupno smanjenje kriterija koje je uzorkovala ta značajka. Kriterij je *Gini* nečistoća, a ona mjeri nečistoću čvora u stablu odluke, s većom težinom za najvažnije značajke, zbog čega je *Gini* važnost poznata kao ukupno smanjenje nečistoće

čvora. *Gini* važnost proširuje ovaj koncept kako bi procijenila doprinos svake značajke kroz više stabala. Prilikom izgradnje slučajne šume, algoritam konstruira skup stabala odluke ponovljenim uzorkovanjem skupa podataka i stvaranjem raznolikih podskupova. Kod slučajne šume za svako stablo *Gini* važnost kvantificira smanjenje *Gini* nečistoće postignuto dijeljenjem čvorova na temelju određene značajke. Dakle, ako neka značajka može uzrokovati jasne podjele u stablu i doprinosi čišćoj podjeli podataka, ta će značajka imati veću važnost [19].

Permutacijska važnost procjenjuje važnost svake značajke nezavisno te se smatra pouzdanijom i robusnijom mjerom. Evaluira se utjecaj permutiranja pojedine značajke na predviđanje te se tako računa važnost. Ova metoda uključuje nasumično mijenjanje vrijednosti neke značajke, nakon čega slijedi mjerenje pada u izvedbi modela. Potom se uspoređuju izvorne i permutirane vrijednosti značajke kako bi se precizno izmjerila važnost te značajke. Što je pogoršanje u izvedbi modela veće, to je značajka važnija [19].

Metoda za računanje permutacijske važnosti je opisana sljedećim koracima [20]:

- ulazi: trenirani model m, testni skup podataka D
- izračunaj referentnu točnost s modela m na podacima D (u slučaju regresije, R^2)
- **z**a svaku značajku j (stupac iz skupa D):
  - za svako ponavljanje k u $\{1, \ldots, K\}$:
    - nasumično mijenjaj vrijednosti značajke j iz skupa D kako bi se stvorila nova verzija skupa podataka $\tilde{D}_{k,j}$
    - izračunaj rezultat $s_{k,j}$ modela m na podacima s permutiranim značajkama $\tilde{D}_{k,j}$
  - izračunaj važnost i_j za značajku f_j dobivenu formulom 4.1:

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{k,j} \tag{4.1}$$

5. Rezultati i rasprava

5.1. Mjere točnosti

Nakon što su objašnjene izvedbe modela, predstavit će se njihovi rezultati. Za evaluaciju izvedbi regresijskih modela koristit će se tri mjere: srednja apsolutna greška (engl. *Mean Absolute Error*, MAE), korijen srednje kvadratne greške (engl. *Root Mean Squared Error*, RMSE) i koeficijent determinacije (engl. *Coefficient of Determination*, R^2).

Srednja apsolutna greška se računa kao srednja vrijednost apsolutne razlike stvarnih i predviđenih vrijednosti. Izraz se dobiva pomoću formule 5.1

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |(y_i - \hat{y}_i)| \tag{5.1}$$

gdje je N ukupan broj podataka, y_i je predviđena vrijednost, a $\hat{y}_i$ srednja vrijednost za y. Razlika između te dvije vrijednosti poznata je kao *rezidual*, stoga se može reći da MAE mjeri srednju vrijednost reziduala u skupu podataka [21].

Srednja apsolutna greška je funkcija troška koja se može predstaviti kao korijen srednje kvadratne greške (engl. *Mean Squared Error*, MSE). RMSE se smatra boljom mjerom od MSE jer normalizira podatke kako bi bili na istoj ljestvici kao i ciljana značajka. RMSE mjeri standardnu devijaciju reziduala te se računa formulom 5.2 [21].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{5.2}$$

Koeficijent determinacija objašnjava koji omjer varijance u izlaznoj značajki je objaš-

njiv ulaznim značajkama odnosno modelom. R^2 nije osjetljiv na različite ljestvice te može poprimiti vrijednosti između *0* i *1*. R^2 se računa pomoću formule 5.3 [21].

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{5.3}$$

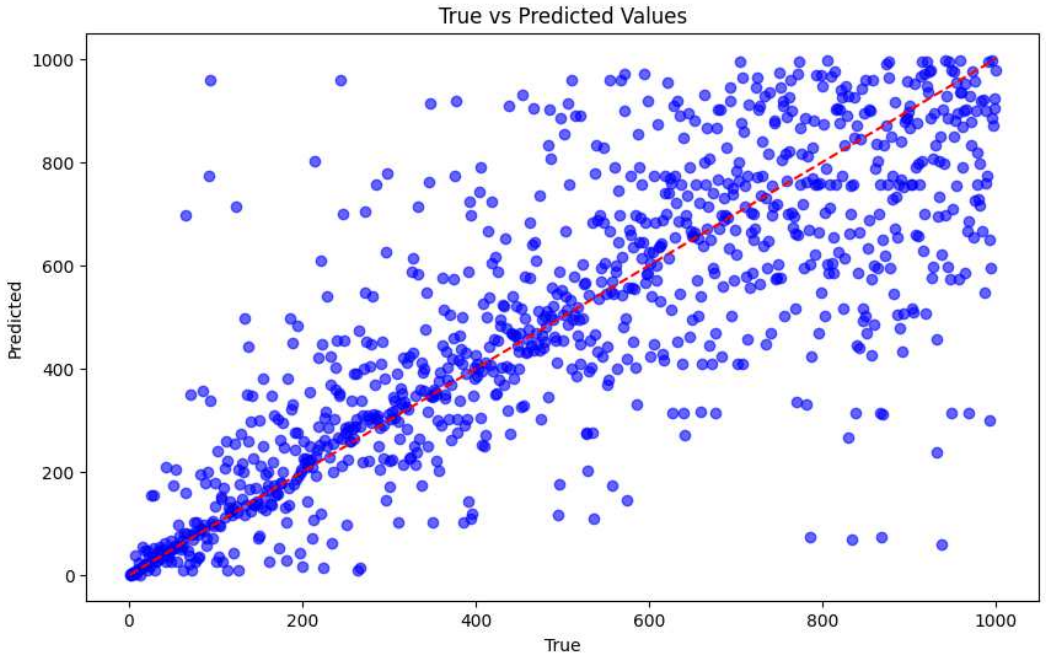
Kada se mjeri kvaliteta modela, poželjno je da vrijednosti mjera MAE i RMSE budu što manje, a vrijednost R^2 što veća.

Treba napomenuti da su se izlazi modela zaokružili na cijeli broj zato što rang predstavlja diskretne vrijednosti pa su se mjere točnosti računale pomoću tih vrijednosti.

5.2. Rezultati implementiranih modela

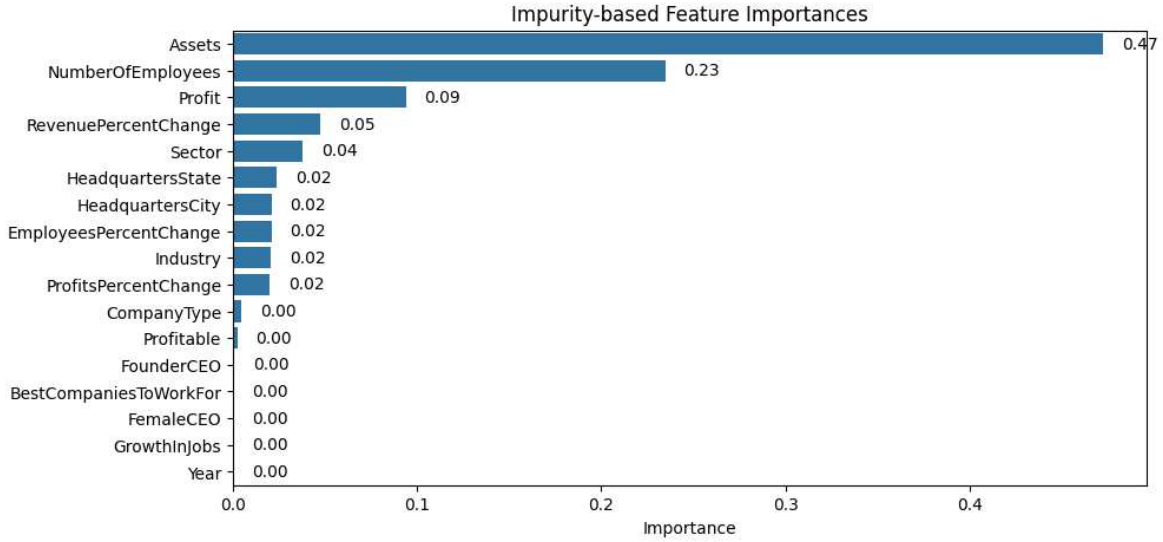
5.2.1. Stablo odluke

Nakon implementacije stabla odluke sa zadanim parametrima u *scikit-learn* biblioteci dobivamo sljedeće rezultate: 123,798 za MAE, 182,946 za RMSE te 0,598 za R^2. Također smo rezultat prikazali grafom na Slici 5.1., gdje *x-os* označava stvarnu vrijednost, a *y-os* označava predviđenu vrijednost ranga. Što je točka na dijagramu bliža pravcu $y = x$, to je predviđena vrijednost bliža stvarnoj vrijednosti.

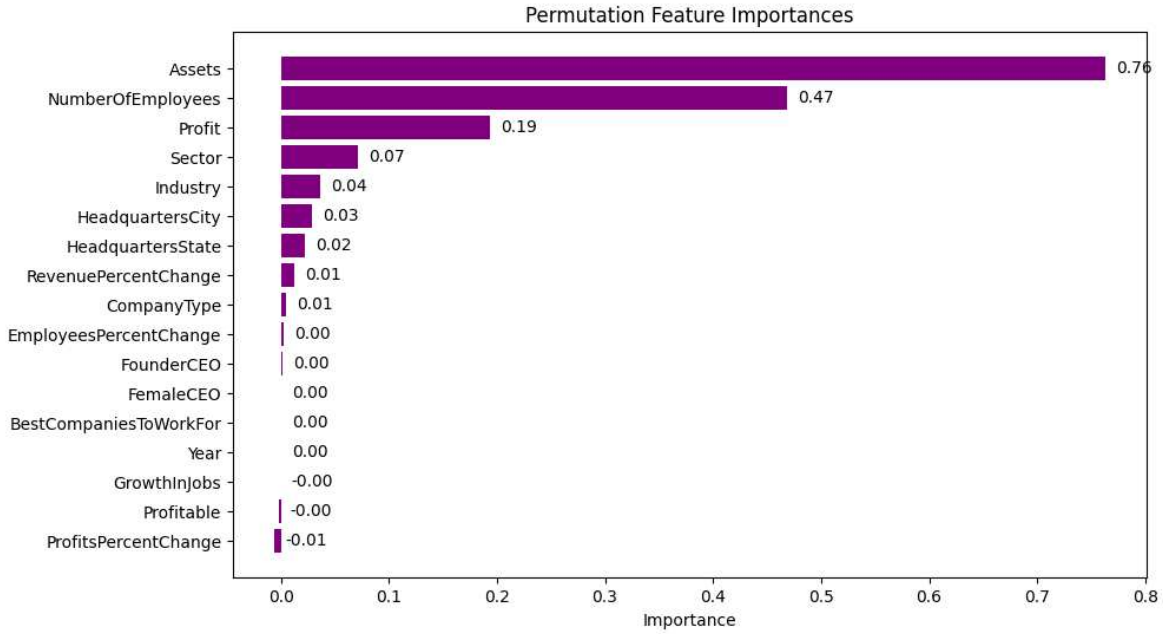


Slika 5.1. Stvarne i predviđene vrijednosti za stablo odluke

Zatim smo nakon identifikacije najvažnijih značajkih pomoću dviju metoda dobili rezultate prikazane na Slici 5.2. za važnost temeljenu na nečistoći i Slici 5.3. za permutacijsku važnost.



Slika 5.2. Važnost značajki temeljena na nečistoći za stablo odluke



Slika 5.3. Permutacijska važnost značajki za stablo odluke

Vidimo da su dvije metode dale slične rezultate, ali ne identične. Pri vrhu ljestvice se na oba prikaza nalaze sljedeće značajke: *Assets, NumberOfEmployees, Profit, Sector*. Osim njih, preostale istaknute značajke su i *RevenuePercentChange*, *HeadquartersState*, *Industry* i *HeadquartersCity*.

Nakon što smo identificirali značajke koje su najviše utjecale na predviđanje, proveli smo optimizaciju hiperparametetara na modelu s reduciranim brojem ulaznih značajki, točnije prethodno navedenih osam značajki. Vrijednosti hiperparametara su u nastavku:

- *min_samples_split = 4*
- *min_samples_leaf = 10*
- *max_features = None*
- *max_depth = 21*

Zatim je ponovno evaluiran model na testnom skupu. Usporedba rezultata prije i poslije opimizacije dana je Tablicom 5.1.

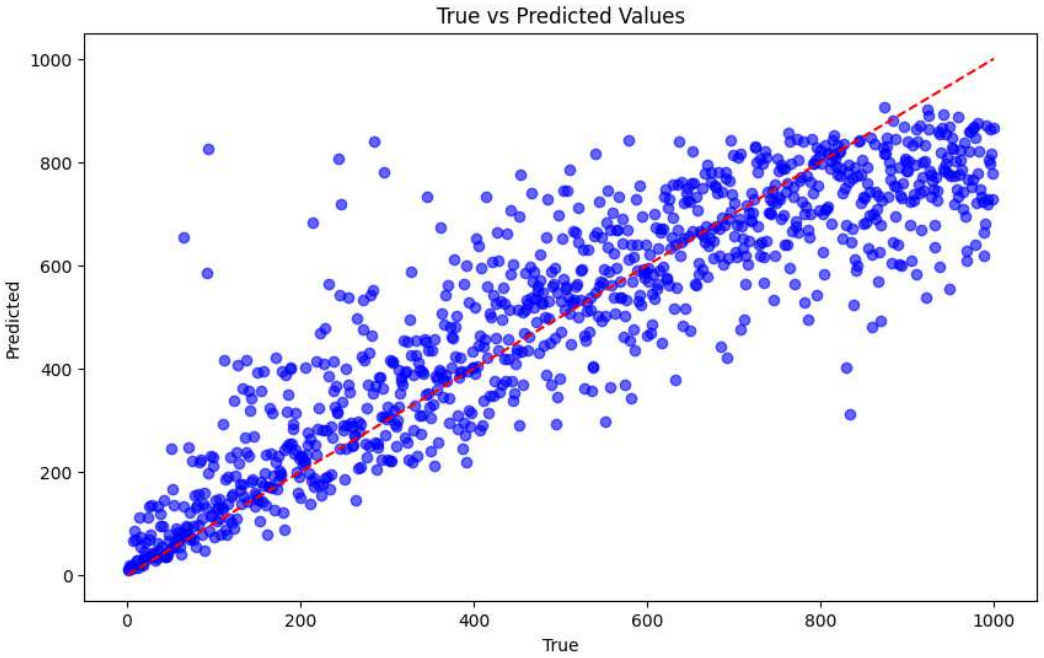
Tablica 5.1. Rezultati modela stabla odluke prije i poslije optimizacije

Mjera točnosti	**Prije**	**Poslije**
MAE	123,798	111,835
RMSE	182,946	12,279
R^2	0,598	0,727

Iz prikaza se može zaključiti da su se rezultati znatno poboljšali nakon redukcije značajki i optimizacije hiperparametara.

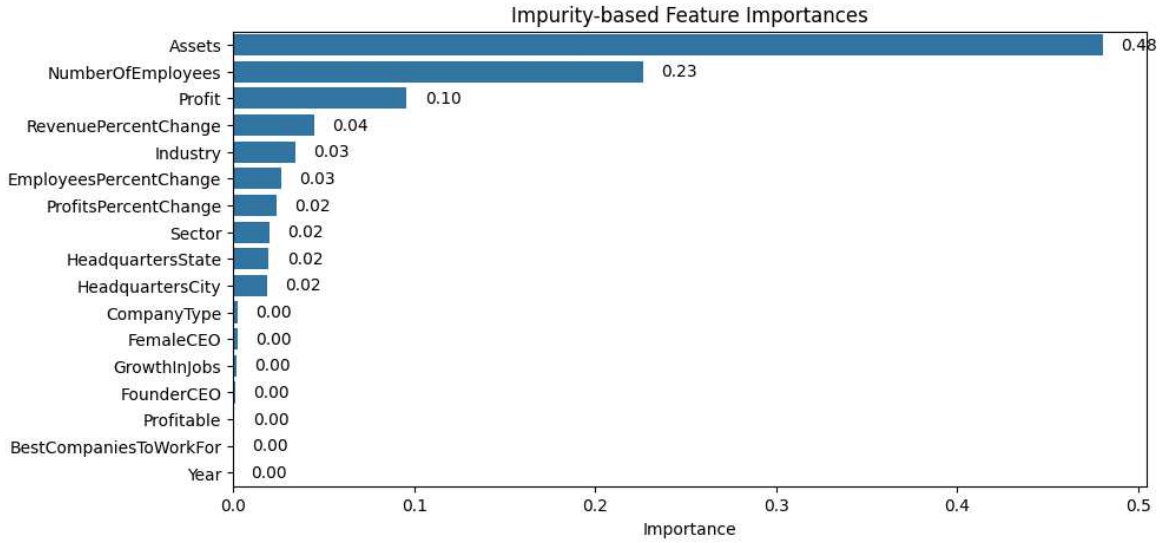
5.2.2. Slučajna šuma

Sljedeći implementirani algoritam je slučajna šuma. Model je najprije izveden sa zadanim parametrima te su dobiveni sljedeći rezultati: MAE od 94,393, RMSE od 130,849 te R^2 od 0,794. Razlika između predviđenih i stvarnih vrijednosti prikazana je grafom na Slici 5.4.

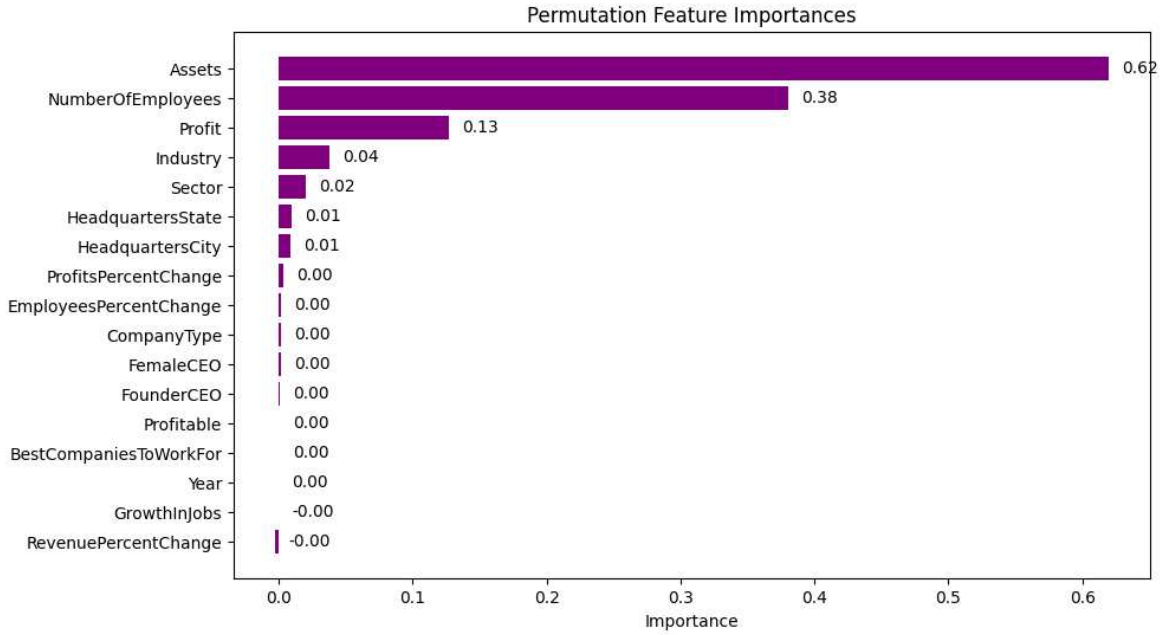


Slika 5.4. Stvarne i predviđene vrijednosti za slučajnu šumu

Potom su se pomoću dviju metoda identificirale značajke koje najviše doprinose predviđanju. Rezultati su prikazani na Slici 5.5. i Slici 5.6.



Slika 5.5. Važnost značajki temeljena na nečistoći za slučajnu šumu



Slika 5.6. Permutacijska važnost značajki za slučajnu šumu

Među najvažnijim značajkama kod obje tehnike istaknule su se sljedeće značajke: *Assets, NumberOfEmployees, Profit, Sector*. Osim njih zadržat ćemo *RevenuePercentChange*, *HeadquartersState*, *Industry* i *HeadquartersCity*. Potom slijedi optimizacija modela s ulaznim značajkama koje su prethodno navedene. Dobiveni hiperparametri nakon optimizacije su:

- $min\_samples\_split = 6$
- $max\_depth = 36$
- $min\_samples\_leaf = 1$
- $max\_features = log2$

Uspoređujemo rezultate prije i poslije optimizacije u Tablici 5.2.

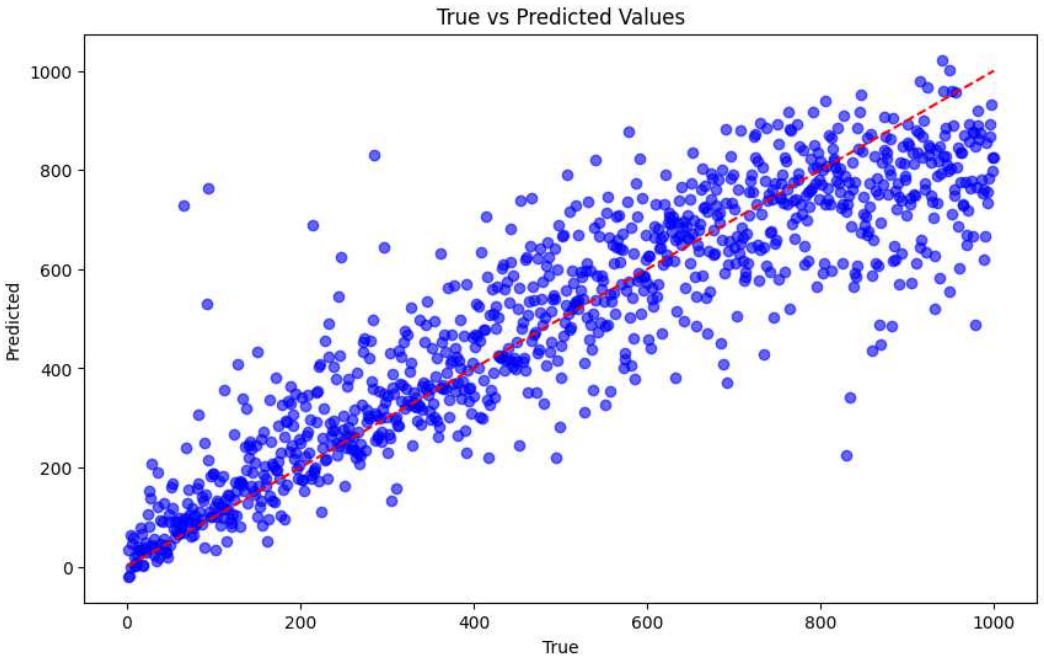
Tablica 5.2. Rezultati modela slučajne šume prije i poslije optimizacije

Mjera točnosti	**Prije**	**Poslije**
MAE	94,393	94,772
RMSE	130,849	11,393
R^2	0,794	0,797

Iz prikaza se da zaključiti da je nakon redukcije značajki i pronalaska optimalnijih hiperparametara model dao bolje rezultate u RMSE i R^2, ali ne i u MAE.

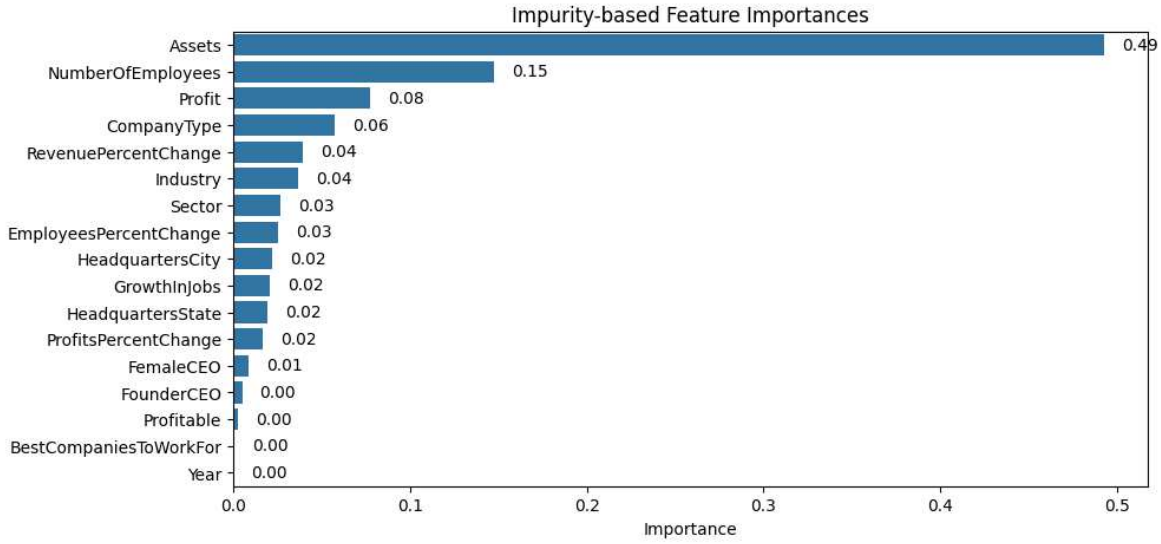
5.2.3. XGBoost

Model XGBoost prvotno je također implementiran sa zadanim parametrima te su dobiveni sljedeći rezultati: 89,621 za MAE, 125,977 za RMSE te 0,809 za R^2. Dijagramom raspršenja prikazan je odnos između predviđenih i stvarnih vrijednosti na Slici 5.1.

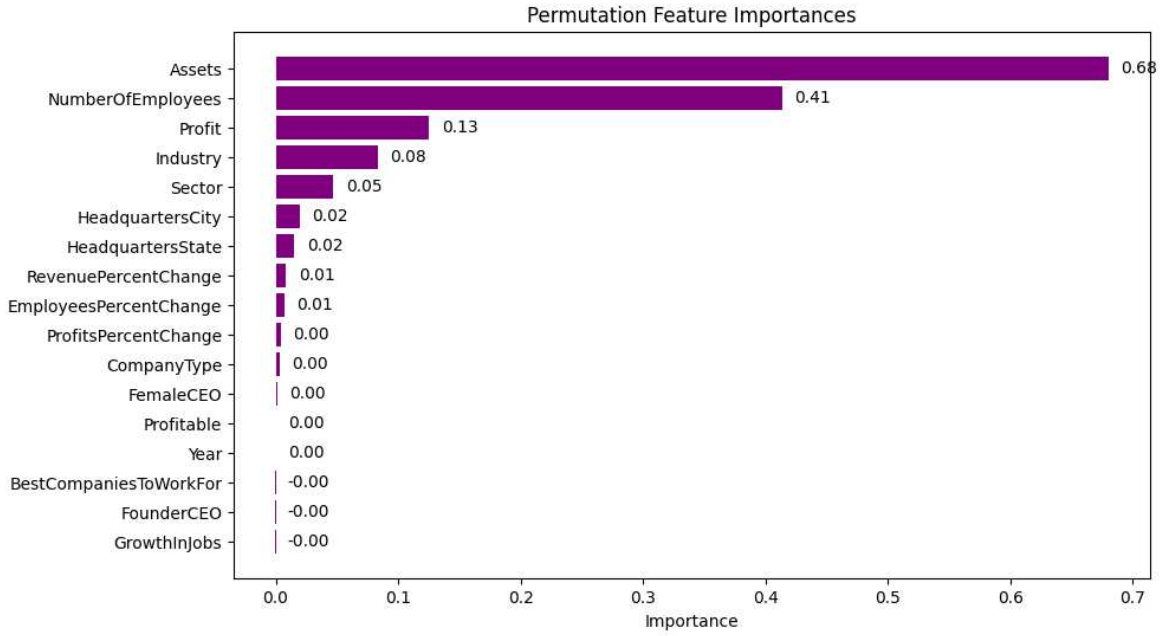


Slika 5.7. Stvarne i predviđene vrijednosti za XGBoost

Koristeći važnost temeljenu na nečistoći i permutacijsku važnost identificirali smo ključne značajke za predviđanje. Rezultati su prikazani na Slici 5.8. za važnost temeljenu na nečistoći i Slici 5.9. za permutacijsku važnost.



Slika 5.8. Važnost značajki temeljena na nečistoći za XGBoost



Slika 5.9. Permutacijska važnost značajki za XGBoost

Dobivaju se vrlo slični rezultati kao i za modele stabla odluke i slučajne šume. Pri vrhu ljestvica se na oba prikaza nalaze značajke: *Assets, NumberOfEmployees, Profit, Industry, Sector*. Osim njih, preostale istaknute značajke su i *RevenuePercentChange*, *HeadquartersState*, *HeadquartersCity* te *CompanyType*.

Nakon toga je provedena optimizacija hiperparametetara na modelu s reduciranim brojem ulaznih značajki, gdje smo ostavili jednakih osam značajki kao u prethodnim

modelima. Rezultati identifikacije najvažnijih značajki su vrlo slični za sva tri modela, stoga ćemo ostaviti iste značajke za reducirani ulaz kako bismo mogli uspoređivati izvedbe triju modela. Hiperparametri nakon optimizacije su iznosili:

- *learning_rate = 0,099*
- *max_depth = 4*
- *n_estimators = 187*
- *subsample = 0,813*

Poslije toga je optimizirani model ponovno evaluiran, a rezultate možemo usporediti s prijašnjim rezultatima u Tablici 5.3.

Tablica 5.3. Rezultati modela XGBoost prije i poslije optimizacije

Mjera točnosti	**Prije**	**Poslije**
MAE	89,621	84,823
RMSE	125,977	10,901
R^2	0,809	0,830

Može se zaključiti da je XGBoost dosad dao najbolje rezultate i prije i poslije optimizacije. Također, rezultati su se znatno poboljšali nakon optimizacije, što je evidentno uspoređujući sve tri mjere.

5.3. Rasprava

Nakon što su opisane izvedbe triju modela te predstavljeni rezultati njihovih izvedbi, slijedi osvrt na dobiveno. XGBoost se može izdvojiti kao model koji je dao najbolje rezultate i prije i poslije optimizacije. Postigao je R^2 mjeru od čak 0,809 prije optimizacije, što je nešto veća razlika od modela slučajne šume koji je postigao 0,794 i znatno veći rezultat u odnosu na stablo odluke koje je dalo R^2 mjeru 0,598.

Sljedeći korak bio je identifikacija najvažnijih značajki za predviđanje modela s ciljem unapređenja izvedbe modela. Dobiveno je da su za sva tri modela identificirane sljedeće značajke: *Assets, NumberOfEmployees, Profit, Sector*. Istakle su se i značajke: *RevenuePercentChange, Industry, HeadquartersCity i HeadquartersState*. U skladu s očekivanjem je dobiveno da će na uspjeh poduzeća uvelike utjecati imovina tog poduzeća,

broj zaposlenika i dobit koju ustvaruje. Došlo se i do zanimljivog zaključka da sektor i industrija također utječu na rang poduzeća, a time implicitno i na ostvarene prihode. Nešto manji značaj je imala lokacija poduzeća, no unatoč tome utjecala je na predviđanja modela, što nije bilo očekivano na početku.

Nakon optimizacije parametara na modelu s reduciranim značajkama dobiveni su bolji rezultati, što govori da su dobivene značajke uistinu važne za model te da treba pažljivo odabrati prostor hiperparametara s kojim se radi. U Tablici 5.4. dana je usporedba rezultata svih modela prije i poslije optimizacije.

Tablica 5.4. Rezultati modela prije i poslije optimizacije

	DT		**RF**		**XGBoost**	
Mjera	**Prije**	**Poslije**	**Prije**	**Poslije**	**Prije**	**Poslije**
MAE	123,798	111,835	94,393	94,772	89,621	84,823
RMSE	182,946	12,279	130,849	11,393	125,977	10,901
R^2	0,598	0,727	0,794	0,797	0,809	0,830

Uspoređujući rezultate, jasno je da XGBoost daje najbolje rezultate u svakom pogledu. Rezultati nakon optimizacije su minimizirani za MAE (84,823) i RMSE (10,901) i maksimizirani za R^2 (0,830).

Osim korištenih modela treba uzeti u obzir kvalitetu i kvantitetu korištenih podataka. Modeli su se trenirali na povijesnim podacima samo za 2023. godinu, što se može unaprijediti u daljnjim istraživanjima uključujući podatke iz više prethodnih godina kako bi model imao više ulaznih podataka. Očekuje se da model daje bolje rezultate što više podataka ima. U ovom slučaju ograničenje je bilo dostupnost podataka na internetu, što bi se moglo promijeniti u budućnosti.

Također, osim korištenih modela stabala za predviđanje se mogu koristiti te usporediti i drugi regresijski algoritmi kao što su: K-najbližih susjeda, regresor stroja potpornih vektora (engl. *Support Vector Machine Regressor*) i sl.

Iako su korišteni modeli i analizirane značajke pružili vrijedne uvide, treba naglasiti da na uspješnost poduzeća utječu i brojni vanjski faktori koji nisu izravno obuhvaćeni u ovom istraživanju. Globalna ekonomska situacija, makroekonomski pokazatelji poput inflacije, kamatnih stopa i bruto domaćeg proizvoda (BDP), kao i geopolitičke okolnosti

mogu značajno oblikovati poslovanje poduzeća.

Primjerice, pandemija COVID-19 pokazala je kako nepredvidivi globalni događaji mogu imati iznimno velik utjecaj na profitabilnost i stabilnost poduzeća, čak i u sektorima koji tradicionalno bilježe stabilne prihode. Nadalje, klimatske promjene i zakonodavni pritisci usmjereni prema održivosti sve više utječu na strategije poduzeća, osobito u energetskom i proizvodnom sektoru.

Osim toga, promjene u tehnološkim trendovima, potrošačkim preferencijama te konkurenciji na tržištu također igraju ključnu ulogu u oblikovanju uspješnosti. Ove dinamične okolnosti naglašavaju potrebu za integriranjem šireg spektra značajki i izvora podataka u budućim istraživanjima, kako bi se obuhvatili i vanjski faktori koji utječu na poslovanje.

6. Zaključak

Ovaj diplomski rad istraživao je mogućnost predviđanja uspješnosti poduzeća korištenjem različitih modela strojnog učenja te identifikacijom značajki koje najviše doprinose preciznosti predviđanja. Cilj istraživanja bio je implementirati i evaluirati modele za predviđanje ranga poduzeća na temelju javno dostupnih podataka te analizirati utjecaj odabranih značajki na rezultate.

Provedene analize i implementacije potvrdile su da modeli poput XGBoosta, slučajne šume i stabla odluke mogu učinkovito predvidjeti rang poduzeća. Među njima, XGBoost se izdvojio kao najuspješniji model s najboljim rezultatima točnosti, pokazujući značajnu robusnost u odnosu na stršeće vrijednosti i složenost podataka. Također, optimizacija hiperparametara i redukcija značajki dodatno su poboljšale performanse modela.

Identifikacija ključnih značajki otkrila je da imovina poduzeća, broj zaposlenika, dobit i sektor poslovanja imaju najvažniji utjecaj na predviđanja. Ovi rezultati sugeriraju da uspjeh poduzeća nije samo financijska, već i strateška tema, pri čemu različiti sektori poslovanja igraju važnu ulogu u pozicioniranju poduzeća na tržištu.

Zaključno, rad pokazuje da se metodologija strojnog učenja može učinkovito koristiti za analizu poslovnih podataka i donošenje informiranih poslovnih odluka. Rezultati ovog rada pružaju smjernice za daljnje istraživanje, uključujući primjenu dodatnih modela i proširenje analize na globalne podatke ili podatke iz različitih industrija.

Literatura

[1] Y. Gao, Y. Luo, i C. Pan, "Machine Learning Prediction of Companies ' Business Success", u *CS229: Machine Learning, Fall 2018, Stanford University, CA*, 2018. [Mrežno]. Adresa: https://api.semanticscholar.org/CorpusID:86862672

[2] A. Ali, "fortune1000-ml-prediction", https://github.com/ArshadAliDS/fortune1000-ml-prediction, 2024., pristupljeno 20. listopada 2024.

[3] "What Is Machine Learning (ML)?" https://www.ibm.com/think/topics/machine-learning, 2021., pristupljeno 23. listopada 2024.

[4] "Decision Tree", https://www.geeksforgeeks.org/decision-tree/, 2017., pristupljeno 1. studenog 2024.

[5] "Machine Learning Random Forest Algorithm", https://www.javatpoint.com/machine-learning-random-forest-algorithm, 2024., pristupljeno 4. studenog 2024.

[6] "A Visual Comparison between the Complexity of Decision Trees and Random Forests", https://commons.wikimedia.org/wiki/File:Decision_Tree_vs._Random_Forest.png, 2020., pristupljeno 4. studenog 2024.

[7] "What Is XGBoost?" https://www.nvidia.com/en-us/glossary/xgboost/, 2024., pristupljeno 10. studenog 2024.

[8] "K-Nearest Neighbor (KNN) Algorithm", https://www.geeksforgeeks.org/k-nearest-neighbours/, 2017., pristupljeno 10. studenog 2024.

[9] "Schematic View of K-Nearest Neighbors Methods", https://commons.wikimedia.org/wiki/File:K-nearest_Neighbors.png, 2015., pristupljeno 10. studenog 2024.

[10] J. Chen, “Fortune 1000: Annual List of Largest American Companies”, https://www.investopedia.com/terms/f/fortune-1000.asp, pristupljeno 1. prosinca 2024.

[11] “Ml | handling missing values”, https://www.geeksforgeeks.org/ml-handling-missing-values/, 2018., pristupljeno 4. prosinca 2024.

[12] “Handling Missing Data with KNN Imputer”, https://www.geeksforgeeks.org/handling-missing-data-with-knn-imputer/, 2024., pristupljeno 4. prosinca 2024.

[13] A. A. Masud, “Correlation Matrix: What Is It, How It Works Examples.” https://www.questionpro.com/blog/correlation-matrix/, 2023., pristupljeno 10. prosinca 2024.

[14] B. Wohlwend, “Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning”, https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948, 2023., pristupljeno 15. prosinca 2024.

[15] A. Chawla, “Grid Search vs. Random Search vs. Bayesian Optimization”, https://blog.dailydoseofds.com/p/grid-search-vs-random-search-vs-bayesian, 2024., pristupljeno 15. prosinca 2024.

[16] “DecisionTreeRegressor”, https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html, pristupljeno 7. siječnja 2025.

[17] “RandomForestRegressor”, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, pristupljeno 7. siječnja 2025.

[18] B. Priya, “Hyperparameter Tuning: GridSearchCV and RandomizedSearchCV, Explained”, https://www.kdnuggets.com/hyperparameter-tuning-gridsearchcv-and-randomizedsearchcv-explained, 2023., pristupljeno 15. siječnja 2025.

[19] “Feature Importance with Random Forests”, https://www.geeksforgeeks.org/feature-importance-with-random-forests/, 2017., pristupljeno 18. siječnja 2025.

[20] “Permutation Feature Importance”, https://scikit-learn.org/1.5/modules/permutation_importance.html, pristupljeno: 21. siječnja 2025.

[21] A. Chugh, “MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?” https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e, pristupljeno 21. siječnja 2025.

Sažetak

IDENTIFIKACIJA NAJVAŽNIJIH ZNAČAJKI U MODELU PREDVIĐANJA USPJEŠNOSTI PODUZEĆA

Ivana Kuliš

Rad istražuje mogućnosti primjene strojnog učenja u predviđanju uspješnosti poduzeća temeljenog na javno dostupnim podacima Fortune 1000. Cilj rada je bio identificirati ključne značajke koje najviše doprinose točnosti predviđanja ranga poduzeća i analizirati njihov utjecaj. Koristeći algoritme stabla odluke, slučajne šume i XGBoost, implementirani su modeli za predviđanje, a potom optimizirani s ciljem poboljšanja njihove izvedbe. Provedena je eksploratorna analiza podataka kako bi se razumjeli odnosi između značajki poput prihoda, dobiti, broja zaposlenika i sektora poslovanja.

Rezultati pokazuju da XGBoost postiže najbolje rezultate, s koeficijentom determinacije (R^2) od 0,830 nakon optimizacije. Najvažnije značajke identificirane u svim modelima uključuju imovinu poduzeća, broj zaposlenika, dobit i sektor poslovanja, dok su lokacija poduzeća i promjene prihoda također imale mjerljiv utjecaj. Rad zaključuje da integracija strojnog učenja i analize podataka omogućava bolje razumijevanje čimbenika koji utječu na poslovni uspjeh te predlaže daljnje proširenje istraživanja na globalne podatke i druge industrije.

Ključne riječi: strojno učenje; Fortune1000; uspjeh poduzeća; stablo odluke; slučajna šuma; XGBoost

Abstract

IDENTIFICATION OF THE MOST IMPORTANT FEATURES IN A COMPANY PERFORMANCE PREDICTION MODEL

Ivana Kuliš

This thesis explores the application of machine learning to predict business success using publicly available Fortune 1000 data. The primary goal was to identify the key features that contribute the most to the accuracy of ranking predictions and to analyze their impact. Decision Tree, Random Forest, and XGBoost algorithms were implemented and optimized to enhance performance. Exploratory data analysis was conducted to understand the relationships between features such as revenue, profit, number of employees, and business sectors.

The results demonstrate that XGBoost achieved the best performance, with a coefficient of determination (R^2) of 0,830 after optimization. The most important features identified across all models include company assets, number of employees, profit, and business sector, while company location and revenue changes also had measurable influence. The study concludes that the integration of machine learning and data analysis enables a better understanding of the factors influencing business success and suggests further research on global datasets and other industries.

Keywords: machine learning; Fortune 1000; business success; decision tree; random forest; XGBoost

Privitak A: Kod

Programski kod i datoteke za učitavanje nalaze se u Github repozitoriju na sljedećoj poveznici: git@github.com:ivanakulis/diplomski.git