FAKULTA INFORMATIKY

A INFORMAČNÝCH TECHNOLÓGIÍ

SLOVENSKÁ TECHNICKÁ UNIVERZITA

Predmet: Objavovanie znalostí

Projekt

# Speed dating experiment

Autori: Bc. Lucia Janíková, Timotej Králik

Cvičiaci: doc. Mgr. Michal Kováč, MSc., PhD.

Cvičenie: Streda 12:00

# Obsah

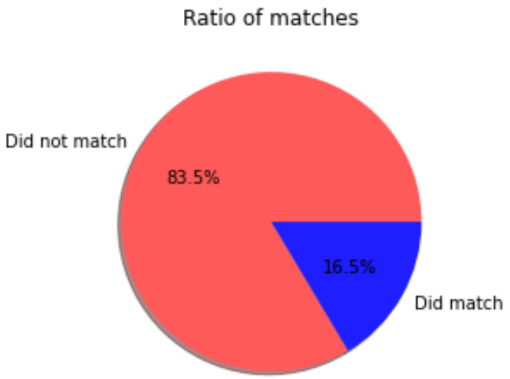
# Dáta

Dataset bol vytvorený na Columbia Business School, konkrétne profesormi Rayom Fismanom a Sheenom Iyengarom pre ich výskumnú prácu *Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment*.
Dáta boli získané od účastníkov niekoľkých experimentov "rýchleho randenia" konaných počas rokov 2002 až 2004.
Každý účastník uviedol základné údaje o sebe ako je vek, vzdelanie, zamestnanie, rasa, tiež ohodnotil svoje vlastnosti (hodnotenie seba v 5 základných atribútoch: inteligencia, zmysel pre humor, atraktivita, ambicióznosť, úprimnosť), uviedol akú dôležitosť dáva jednotlivým záujmom (napr. šport, čítanie, chodenie do divadla alebo múzea, cvičenie jogy,...) či čo očakáva od experimentu.
Počas experimentov účastníci mali 4 minúty na ich "prvé rande" s každým účastníkom opačného pohlavia. Po týchto 4 minútach bola účastníkom položená otázka, či by išli na rande s danou osobou znova. Tiež hodnotili ich potenciálneho partnera v piatich atribútoch: atraktivita, úprimnosť, inteligenciu, zábavnosť a ambicióznosť.
Podobné otázky boli kladené účastníkom opakovane, aj po ukončení samotného rýchleho randenia, na základe čoho by bolo možné sledovať zmeny preferencií na partnera v čase či úspešnosť jednotlivých rande.

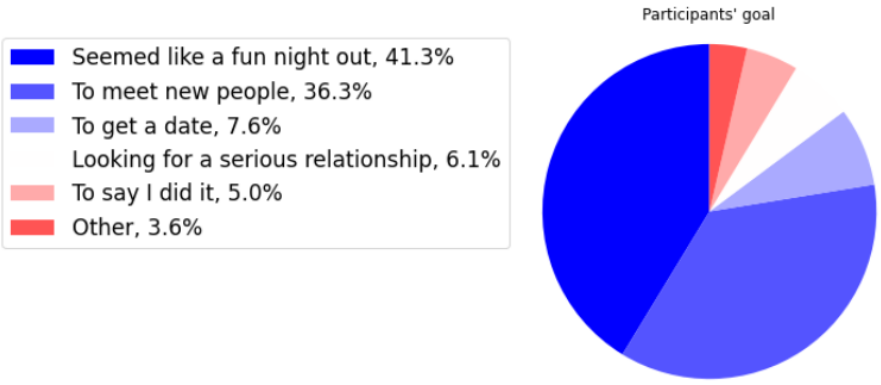
# Exploratívna analýza dát

Dataset obsahuje 8 378 záznamov, pričom každý záznam je tvorený 195 atribútmi. Prakticky všetky atribúty sú kategorické, nakoľko takmer všetky dáta pozostávajú z dotazníkových dát, hodnotení na škále od 1 po 10, tomu sme teda museli prispôsobiť výber aplikovaných techník na analýzu, ale aj celkovú prácu s dátami. Preto sa exploratívna analýza v našom projekte zaoberá skôr skúmaním pomerov v dátach a očakávaných vzťahoch ako klasickými štatistikami typickými pre EDA.

Experimentu sa zúčastnilo 551 účastníkov, z čoho vzniklo 4 189 rande (8 378 záznamov z rande), z čoho iba 16,5% rande skončilo sympatiami na oboch stranách (eng. match).

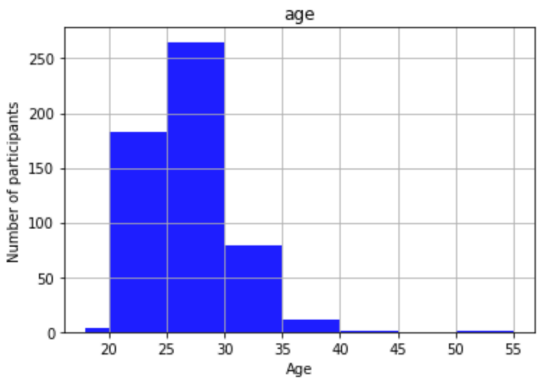


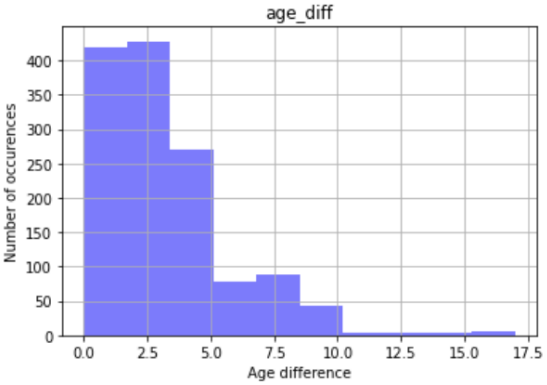

Graf na obrázku nižšie však zobrazuje, že účastníci na experiment neprichádzali primárne s cieľom nájsť si partnera, ale skôr sa zabaviť či spoznať nových ľudí.




## Vek účastníkov

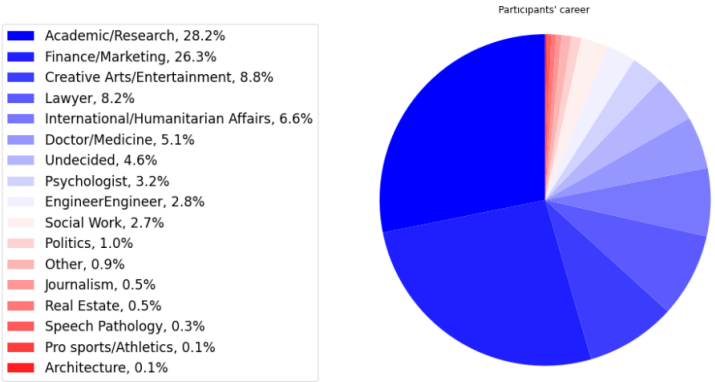
Grafy na obrázkoch nižšie prezentujú vek účastníkov experimentov. Vekové rozpätie účastníkov bolo od 18 po 55 rokov, pričom najväčší počet účastníkov bol vo vekovom rozpätí od 25 do 30 rokov. Väčšina dvojíc, pri ktorých boli sympatie obojstranné boli vo vekovom rozdiele maximálne 3 roky.




## Kariéra

Zúčastnili sa ľudia z rôznych kariérnych oblastí, pričom ako prezentuje graf na obrázku nižšie, najviac účastníkov bolo z akademickej a výskumnej oblasti či z financií a manažmentu, naopak najmenej z oblasti architektúry.

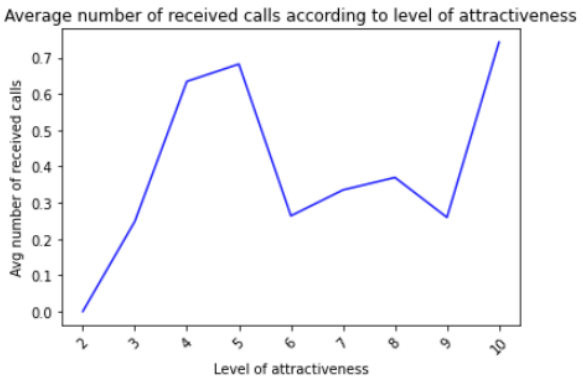


## Volajú účastníci, ktorí chodia častejšie von, častejšie na rande?

Predpokladali by sme, že účastníci, ktorí zvyknú chodiť častejšie von budú aktívnejší v pozvaniach na rande. Dáta však túto teóriu nepotvrdzujú. Ako prezentuje obrázok nižšie, ľudia ktorí sa nachádzajú v priemere frekvencie chodenia von najčastejšie pozývajú opačné pohlavie na rande.


## Obdržia atraktívni ľudia viac telefonátov?

Očakávali by sme, že atraktívnejší účastníci obdržia viac telefonátov od druhého pohlavia ako tí menej atraktívni (škála 1-10, 1=neatraktívny, 10=veľmi atraktívny). Ako prezentuje graf na obrázku nižšie, toto očakávanie sa však nepotvrdilo. Najviac telefonátov obdržali ľudia s atraktivitou na úrovni 5 (prakticky stred intervalu) a potom s očakávanou úrovňou 10 (maximum). Neočakávaný je pokles medzi úrovňou atraktivity 5 a 10.

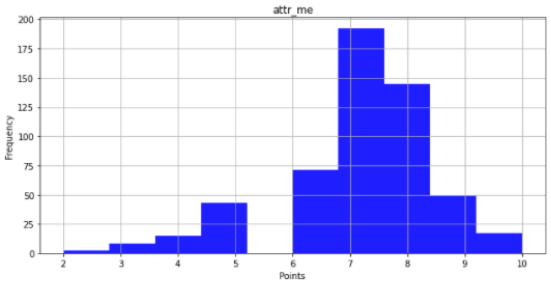


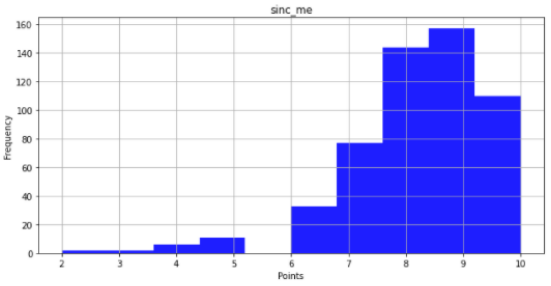

## Hodnotenie svojich vlastných osobnostných charakteristík

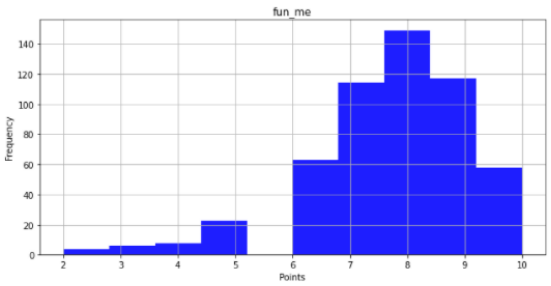
Účastníci hodnotili svojich 5 základných charakteristík na škále 1 (zlé) - 10 (super).
Hodnotené atribúty:

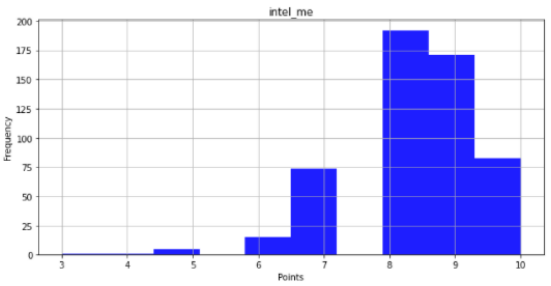
- atr - Atraktivita
- sinc - Úprimnosť
- fun - Zmysel pre humor
- intel - Inteligencia
- amb - Ambicióznosť

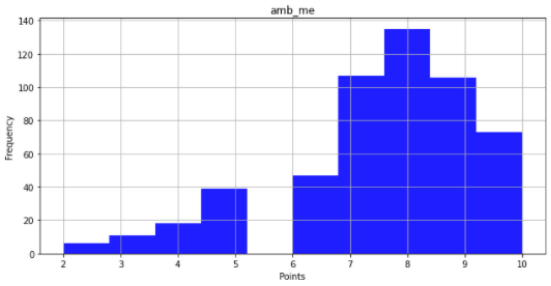
Z dát sme zistili, že účastníci mali najväčšie sebavedomie v inteligencii, ktorá nadobudla priemer na úrovni 8,36.
Z dát je možno vidieť, že naopak, hodnoty majú tendenciu skôr inklinovať ku krajným bodom intervalu a teda účastníci sa vyhýbali stredu intervalu, celkovo je prevaha v druhej polovici intervalu, pri vyšších hodnotách.



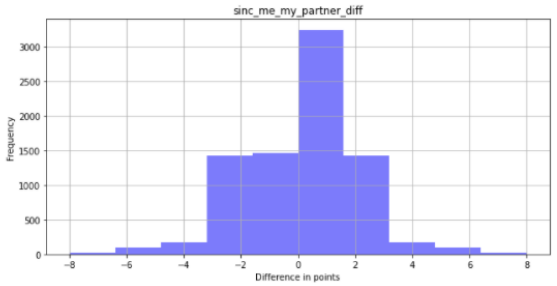



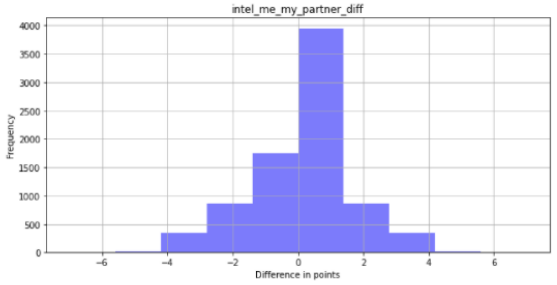

## Hľadajú ľudia partnerov s podobnými vlastnosťami, aké majú oni sami?

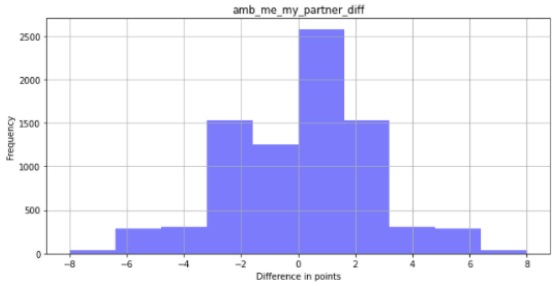
Grafy priložené nižšie prezentujú rozdiel v bodoch medzi účastníkovými vlastnosťami a jeho preferenciami na partnera. Grafy potvrdzujú, že väčšina hľadá partnera s minimálnymi rozdielmi, ľudia preferujú partnerov s rovnakými alebo dokonca trochu lepšími vlastnosťami ako majú oni sami.
Preferencie boli potvrdené aj realitou, graf porovnania vlastností medzi jednotlivými dvojicami, kde boli sympatie obojstranné vyzerá veľmi podobne ako nižšie priložený graf týkajúci sa preferencií.
Škála grafov:

- -9 = nie je to dôležité pre daného človeka, ale malo by to byť dôležité pre jeho partnera
- 0 = žiadny rozdiel v preferenciách
- 9 = je to dôležité pre daného človeka, ale nemalo by to byť dôležité pre jeho partnera



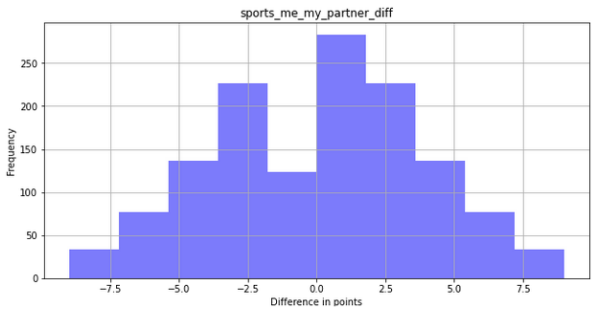


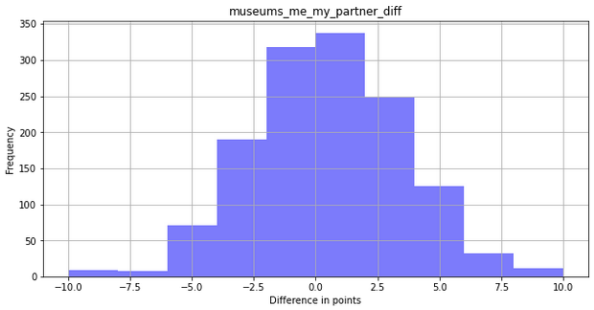




## Akú úlohu zohrávajú spoločné záujmy pri prvom rande?

Sú rovnaké preferencie na hobby pre partnerov dôležité? Pozreli sme sa na rozdiel dôležitosti záujmov účastníka a jeho partnera v prípade, že sympatie boli na oboch stranách. Zo zoznamu záujmov sme pre stručnosť dokumentácie vybrali iba zopár zaujímavých, ktoré sú odprezentované na grafoch nižšie. Z dát zobrazených na grafoch je možné vidieť, že v porovnaní s osobnostnými vlastnosťami partnera, na hobby ľudia nemajú až také prísne požiadavky. Stále platí, že ľudia zväčša hľadajú partnera s podobnou až rovnakou preferenciou na záujmy, no nemožno povedať, že by ich požiadavky boli, aby partner musel mať minimálne takú istú preferenciu na jednotlivé záujmy.
Škála grafov:

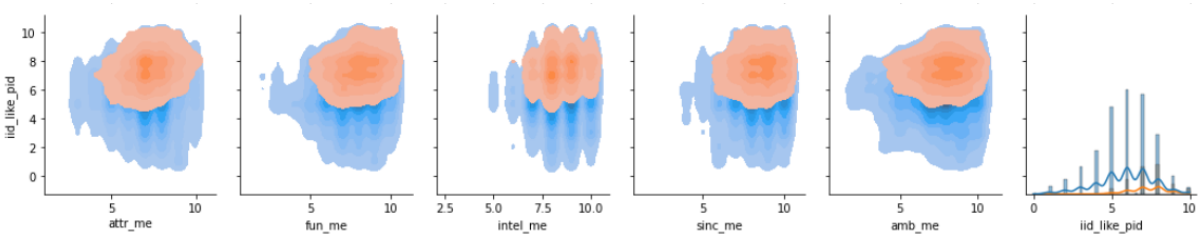
- -9 = nie je to dôležité pre daného človeka, ale malo by to byť dôležité pre jeho partnera
- 0 = žiadny rozdiel v preferenciách
- 9 = je to dôležité pre daného človeka, ale nemalo by to byť dôležité pre jeho partnera



# Korelácie medzi osobnostnými atribútmi a mierou dosiahnutých sympatií

Snažili sme sa vizualizovať závislosť hodnotení v 5 osobnostných atribútoch (os x) a dosiahnutej úrovne sympatií (os y). Na základe párových grafov priložených nižšie vidíme, že ľudia ktorí majú lepšie osobnostné vlastnosti dosahujú väčšie sympatie u opačného pohlavia.

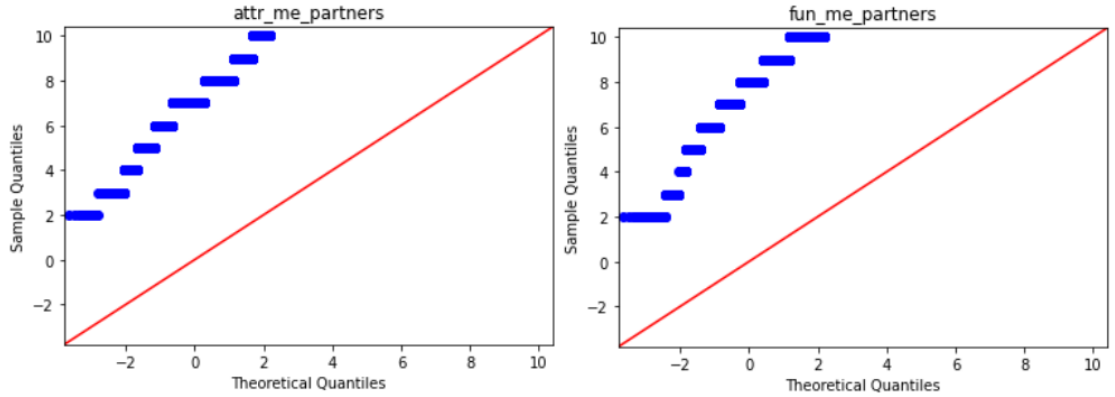


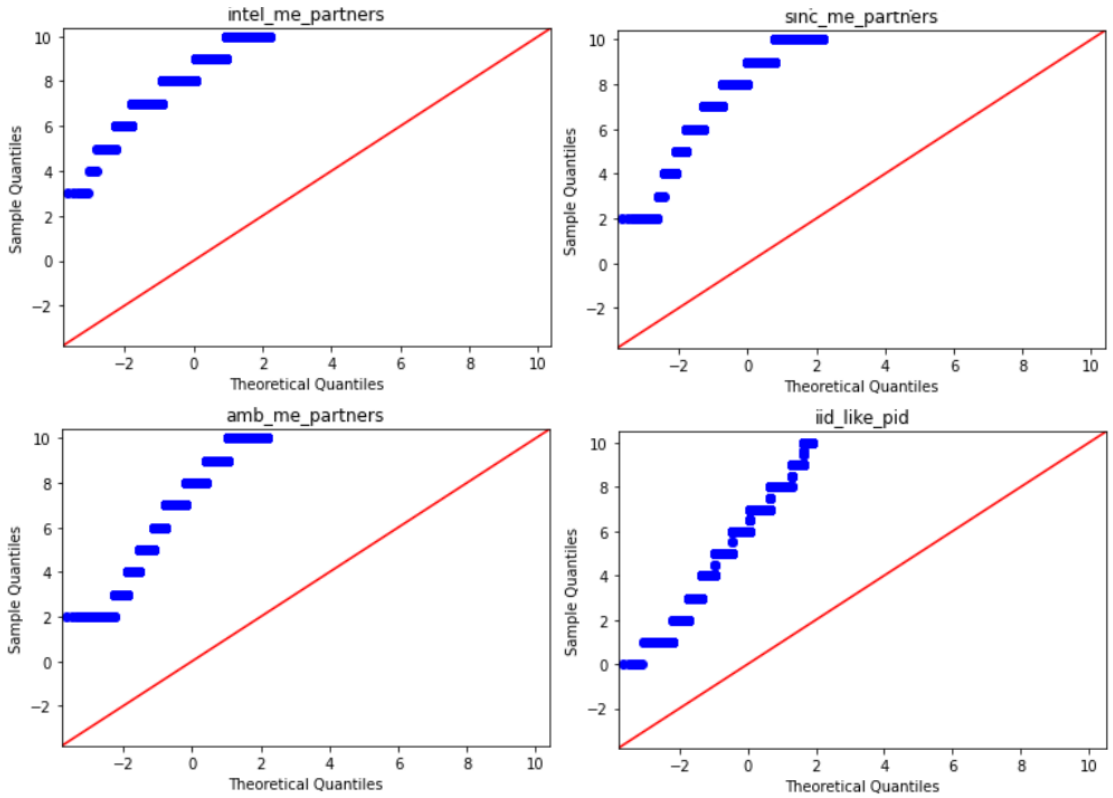

# Šetrenie dát

Väčšina našich dát je v rozmedzí 1-10, z toho vyplýva, že sa v dátach nenachádzajú outliers, ktoré by sme museli predspracovávať. Podobne to je s high leverage points, síce sme ich naozaj len pár v dátach našli (niektorý jedinec nedodržal škálu 1-10). Nakoľko to ale bol veľmi ojedinelý prípad, rozhodli sme sa nevykonať úpravu dát.

Pri atribútoch ohľadom záujmov sme potrebovali rozhodnúť, či sa človek danému záujmu venuje. Toto rozdelenie sme vykonali tak, že ak sa hodnota atribútu nachádza nad celodatasetovým priemerom daného atribútu, človek sa mu venuje.

Mali sme ambíciu používať modely strojového učenia v našich hypotézach a tie mnohokrát implicitne vyžadujú normalitu dát, vykonali sme teda test na základe QQ plotov. QQ ploty pre 5 základných osobnostných atribútov a mieru sympatií medzi partnermi môžeme vidieť nižšie.






Vidíme, že naše dáta nie sú v žiadnom prípade normálne rozložené, k tomuto faktu teda musíme prispôsobiť aj naše riešenia.

# Hypotézy

Na základe analýzy dát sme si pre projekt stanovili 5 hypotéz:

1. Na základe osobnostných vlastností a záujmov je možné predikovať, či sympatie účastníkov budú obojstranné.
2. Na základe vlastností človeka a jeho partnera je možné predikovať ako veľmi sa mu páči a ktoré atribúty majú najväčší vplyv.
3. Na základe vlastností človeka je možné určiť jeho preferenciu na inteligenčnú úroveň jeho partnera (akú inteligenciu majú partneri, ktorých si vybrali).
4. Na základe partnerových preferencií na hobby je možné určiť preferencie na hobby účastníka.
5. Na základe profesie účastníka je možné predikovať profesiu jeho partnera.

## Hypotéza 1

*Na základe osobných vlastností a záujmov je možné predikovať, či sympatie účastníkov budú obojstranné.*

Z dát, ktoré máme k dispozícii nám ako prvá napadla klasická hypotéza, či vieme na základe osobných vlastností dvoch ľudí a ich záujmov predikovať, či sa navzájom budú sebe páčiť a chceli by pokračovať v hlbšom spoznávaní sa.

## Definovanie problému a vybraného modelu

Riešenie tejto hypotézy pozostáva z binárnej klasifikácie, nakoľko máme údaje len o tom ći sa obaja jednotlivci pre seba rozhodli alebo nie, a to na základe
osobných atribútov (každý účastník hodnotil seba v daných atribútoch na škále 1-10):

- Atraktivita
- Úprimnosť
- Zmysel pre humor
- Inteligencia
- Ambicióznosť,

záujmov (každý účastník hodnotil seba v daných atribútoch na škále 1-10):

- Športovanie
- Sledovanie športov
- Trénovanie, bodybuilding
- Jedenie v reštauráciách
- Turistika, kemping
- Hranie hier
- Tancovanie, chodenie do klubov
- Čítanie
- Sledovanie televízie
- Chodenie do divadla
- Filmy
- Chodenie na koncerty
- Hudba
- Nakupovanie
- Joga, meditácia

plus črty, ktoré hovoria o tom, či sú obaja účastníci rande rovnakej rasy a črty obsahujúcu hodnotu korelácie medzi záujmami. Finálne teda chceme do modelu poslať 44 čŕt a predikovať hodnoty 0/1. Na modelovanie tohto typu problému sme sa vybrali smerom frakventistickej štatistiky. Problémom v našom datasete sú čisto kategorické dotazníkové dáta, z tohto dôvodu nemôžeme použiť SVM klasifikátory. Rozhodli sme sa teda pre klasifikátor na princípe k-NN klastrovača. Klastrovač nemá problém s kategorickými dátami a zároveň pomocou scikit-learn je veľmi ľahko použiteľný. Jeho vhodnosť na tento typ problému validujeme aj porovnaním s rozhodovacími stromami.

## Hľadanie optimálneho nastavenia modelu

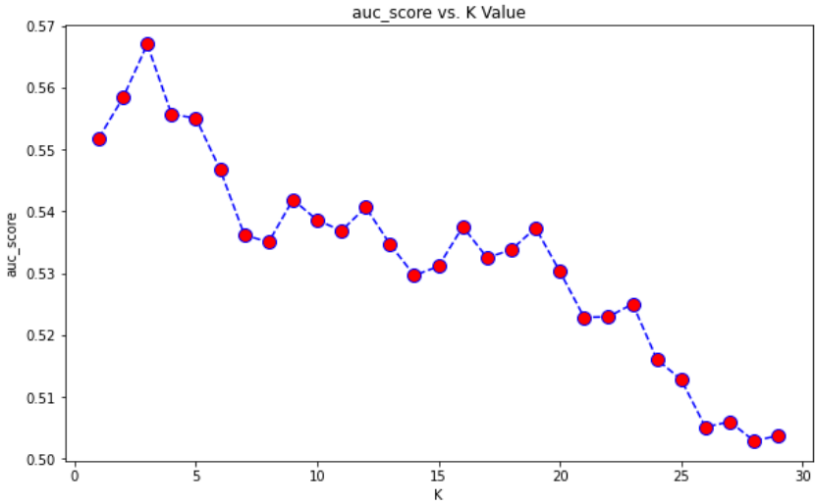
Klastrovanie je silná výpočtová mechanika, ktorá dokáže rozdeliť dáta na skupiny s podobnými vlastnosťami, avšak to má jeden veľký problém, na začiatku nevieme koľko zhlukov sa v našich dátach nachádza. Po uvedomení si faktu, že použijeme klastrovanie, sme teda ako prvé vykonali analýzu, koľko zhlukov pri trénovaní modelu použijeme. Vykonali sme test, kedy sme model natrénovali na počte zhlukov od 1 po 30. Pre každú iteráciu sme si uložili úspešnosť modelu a vybrali najlepší model do finálneho tréningu.

## Metrika validácie modelu

Prichádzame postupne k ďalšiemu podstatného checkpointu nášho postupu riešenia a to je validácia modelu. Ako budeme náš model validovať sa nás spýtalo naše podvedomie a my sme dvojhlasne odpovedali, že ROC AUC krivkami. Vybrali sme si ich kvôli nášmu

nevybalancovanému datasetu. Metrika ako *accuracy* je pre nás samo o sebe nepoužiteľná a na validácia chceme použiť jednu metriku.

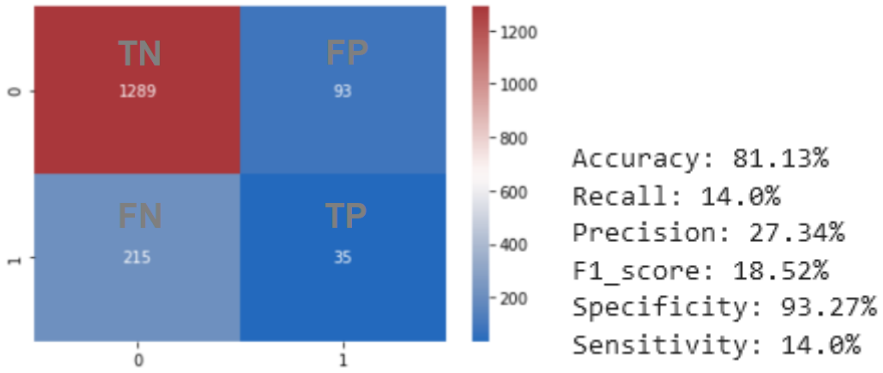
Vykonali sme teda test, kde sme modeli merali ROC AUC hodnotou a prikladáme graf zobrazujúci hodnoty modelov.




Vidíme, že pre naše dáta, je optimálna hodnota počtu klastrov 3. S touto hodnotou sme aj naďalej pracovali a zobrali sme ju do finálneho testovania modelu.

## Porovnávanie viacerých modelov

Porovnali sme štyri verzie modelov. Dva KNN klastrovače a dva rozhodovacie stromy. Z každého druhu modelu sme zobrali dve varianty a to takým spôsobom, že jednu variantu sme trénovali na všetkých trénovacích črtách a druhú len na osobnostných vlastnostiach bez osobných záujmov. Spravili sme tak preto, lebo ako sme videli na začiatku, vieme mať 44 čŕt pri modelovaní. Tento počet je veľký a chceli sme si overiť, či naozaj jednoduchšie modely budú mať horšiu úspešnosť. Z týchto experimentov sme zistil, že KNN klastrovač je mierne úspešnejší v našom probléme ako rozhodovacie stromy a zároveň, že použitie čŕt ohľadom záujmov oboch participantov zvyšuje úspešnosť modelu. Toto zvýšenie však nie je veľmi výrazné, došlo k zlepšeniu modelu o necelého pol percenta v metrike ROC AUC. Uvádzame confusion matrix pre najlepší k-NN klasifikátor, a tiež uvádzame aj základné klasifikačné metriky.



Vidíme, že model preferuje predikciu, že sa nedajú dvaja ľudia spolu do záujmu. Vidíme, že každá štvrtá dvojica, ktorú označíme, že bude mať o seba záujem skutočne má a celkovo odhalíme približne každú siedmu dvojicu, ktorá má skutočne o seba záujem. Keď sme skúsili model optimalizovať vzhľadom na metriku f1_skóre, optimálny počet klastrov sa ukázal na hodnote 1. V tomto prípade sa však znížila hodnota precision pod hranicu 20%. Z pohľadu nášho biznis casu by sme nechceli túto metriku výrazne znižovať, nakoľko odporúčať ľuďom, napriklad na zoznamovacom portáli, že sa k nim hodia partneri, ktorí vôbec nespadajú do ich záujmu, môže spôsobiť odchod používateľov z portálu. Rozhodli sme sa teda preferovať ROC AUC cut off pred F1 skóre.

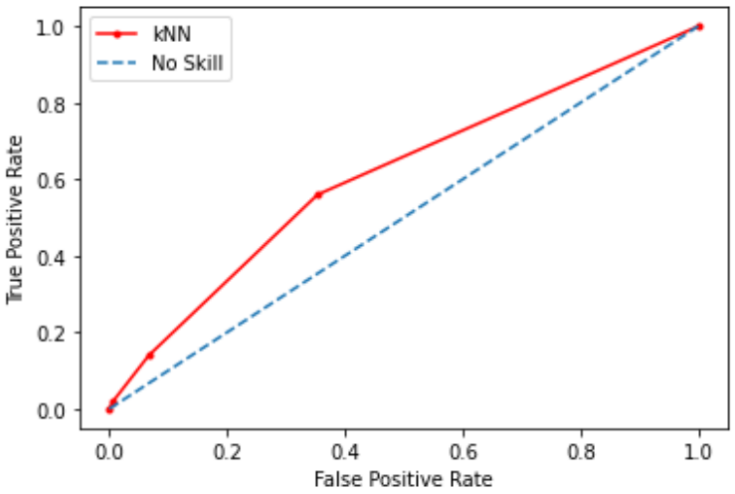
## Krížová validácia

Vstupujeme teda na cieľovú rovinku vo validácií našej hypotézy, kde sme vykonali validácie nášho modelu KNN klastrovača s 3 klastrami cez krížovú validáciu. Túto validáciu sme vykonali tak, že sme model natrénovali 100 krát, pričom sme pre trénovanie zobrali náhodných 80 percent dát a validovali sme na 20-tich. Nad výsledkami tejto validácie sme vykonali súd o našom modeli, či je úspešný. Priemerná hodnota ROC AUC skóre bola 54,80 % pričom štandardná odchýlka bola okolo hodnoty 1 %. Nemôžeme použiť klasickú metriku úspešnosti modelu, nakoľko máme nevybalancovaný dataset, tak sme zvolili ROC AUC hodnotu. Vidíme, že náš model nie je veľmi úspešnejší ako náhoda. Toto sme ale aj očakávali, keď sme vykonávali EDA, lebo dáta sú veľmi rozhodené po priestore a navyše podiel páčiacich sa partnerov medzi sebou bol v menšine.

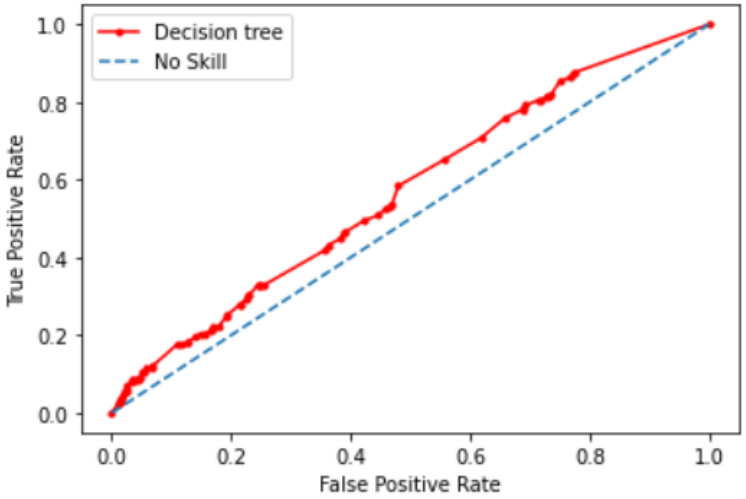
## Štatistická významnosť modelu

Je síce pekné, že sme validovali náš model v rámci nášho biznis casu, ale potrebujeme sa ešte pozrieť na samotnú štatistickú úspešnosť modelu. Využili sme preto Silhouette skóre, aby sme videli metriku, ktorá hovorí o vzájomných závislostiach medzi zhlukmi. Čím je táto hodnota bližšie k nule znamená, že spojitosť jedinca so svojím klastorom je výrazne silnejšia ako spojitosť s klastrami, ku ktorým nepatrí. Táto hodnota nám vyšla len niečo málo väčšia ako 0. Z tohto sme finálne usúdili, že previazanosť medzi klastrami je skoro na takej úrovni ako spojitosť vo vnútri klastra.

Ako dezert podávame ešte pohľad na vývoj našej ROC krivky v tejto klasifikácií.



Očakávateľne sa náš klasifikátor len mierne hýbe nad náhodným klasifikovaním. Ako sme spomínali, testovali sme aj model založený na rozhodovacích stromoch. Dosiahol horšie výsledky ale aj napriek tomu tu uvádzame ROC krivku z najlepšieho modelu založeného na rozhodovacích stromoch.




Vidíme, že ROC krivka, má tvar husľového sláčika, teda úspešnosť modelu je mizerná. Skúšali sme experimentovať aj s jednoduchšími modelmi a chyba sa očakávateľne mierne zväčšovala, avšak keďže sa hýbeme na vrchole hranici chybovosti modelu, už sa chyba nemala veľa priestoru ako zväčšovať.

### Zhodnotenie

Aj keď môžu takto mizerné výsledky vzbudzovať chabý dojem, zamyslime sa nad našou hypotézou so skúsenosťami zo života. Pôvodne sme chceli predikovať či sa láska vzbudí medzi dvoma ľuďmi na základe svojich vyjadrení o osobnosti a záujmov. Jednak tieto dáta sú veľmi zašumené kvôli tomu, že jedna vec je ako sa vnímame my a druhá vec je aká je realita. To, že sa niekto považuje za atraktívneho na 9/10 môže byť úsmevné, hlavne keď je inteligentný na 2/10. Zároveň si myslíme, že dáta nie sú reprezentatívne, nakoľko kultúrne rozdiely veľmi ovplyvňujú spôsoby randenia ľudí a tieto dáta sú zozbierané iba z jednej americkej univerzity, zároveň nepokryli kompletnú vekovú škálu. Druhou vecou je fakt, že láska obsahuje v sebe mystickú hodnotu ktorú nevieme racionálne vyjadriť a keby sme ju s vysokou úspešnosťou dokázali zredukovať len na našu osobnosť a záujmy, výrazne by sme ju zdegradovali a zobrali jej to, čo nás na nej všetkých baví.

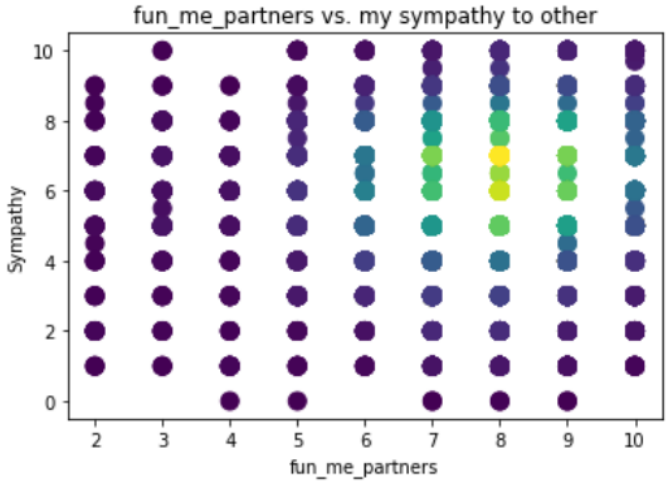
## Hypotéza 2

*Na základe vlastností človeka a jeho partnera je možné predikovať ako veľmi sa mu páči a ktoré atribúty majú najväčší vplyv.*

Existuje nejaký vzťah medzi vlastnosťami človeka a tým, ako sa zapáči druhým ľuďom? Na čom najviac si ľudia zakladajú pri výbere partnera? Je dôležitý vzhľad, zmysel pre humor alebo inteligencia? Z dát, ktoré máme k dispozícii sme sa pokúsili odpovedať na tieto otázky.

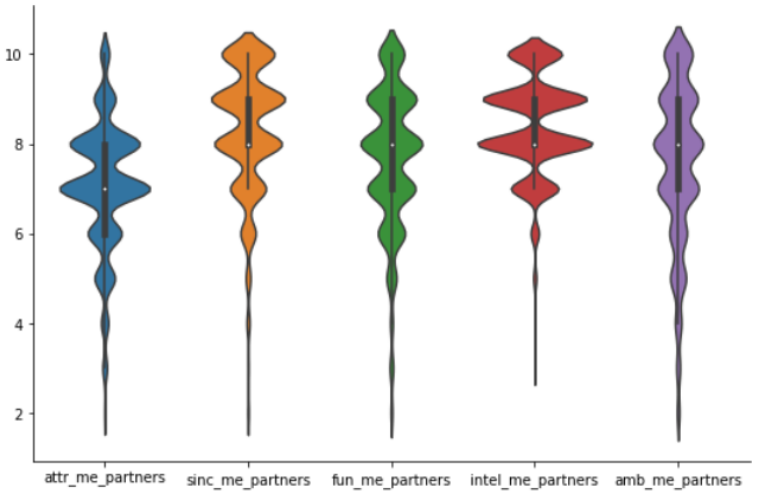
## Definovanie problému a vybraného modelu

Podľa nižšie uvedeného grafu vidíme závislosť medzi zmyslom pre humor účastníka experimentu a miery sympatií získaných od druhého pohlavia (hodnotenie na škále 0-10, 0=najmenej, 10=najviac). Os x prezentuje mieru zmyslu pre humor na škále 1-10 a na osi y je zobrazená miera získaných sympatií. Bledšie farby zobrazujú vysokú intenzitu dosiahnutia danej hodnoty, tmavé naopak, málo prípadov reprezentuje daný bod. Závislosti nie sú úplne priamočiare, je však vidno, že intenzita vyšších sympatií sa zvyšuje s narastajúcim hodnotením zmyslu pre humor, ako prezentuje nasledujúci graf, podobne sa zdá javiť aj atribút atraktivita účastníka.




## Vychýlené hodnoty

O dátach, s ktorými pracujeme sa dá povedať, že prakticky nemajú vychýlené hodnoty, nakoľko sa jedná o dotazníkové dáta, takmer vždy na škále 1-10. Z tohto dôvodu sme nevykonávali žiadnu formu úpravy dát. Graf uvedený na obrázku nižšie reprezentuje vybraných 5 osobnostných vlastností uchádzača (každý uchádzač hodnotil sám seba vo vybraných 5 atribútoch).



## Špecifikácia modelu

Dáta, s ktorými pracujeme nemajú síce Poissonovo rozdelenie (ich stredná hodnota sa štatisticky významne líši od variancie), ale i napriek tomu sme sa rozhodli použiť Poissonovu regresiu na vyhodnotenie našej hypotézy. Nemohli sme použiť klasickú dobre známu logistickú regresiu, nakoľko vieme, že tá dobre funguje na dátach ktoré sú normálne rozdelené a našich kategorických dátach sa takéto rozdelenie nenachádza. Porovnávali sme výsledky aj s Negatívnom binomiálnou regresiou, ktorá nevyžaduje žiadny predpoklad rozdelenia dát, ale Poissonova regresia dosiahla mierne lepšie výsledky, tak sme sa rozhodli pre ňu, aj keď nie je splnený predpoklad rozdelenia dát.
Na obrázku nižšie je možné vidieť výsledky nášho modelu, z ktorých je zrejmé, že najväčší vplyv na výber partnera má jeho zmysel pre humor. Avšak dá sa povedať, že v podstate nie všetky zo zvolených atribútov vypovedajúcich o osobnostných vlastnostiach človeka majú vplyv na sympatie, ktoré vzbudí u opačného pohlavia. Vidíme, že podľa P-value ambicióznosť a úprimnosť neimponujú potencionálnym partnerom. Tieto atribúty sa môžu považovať ako mätúce pre náš model.

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | iid_like_pid | No. Observations: | 6425 |
| Model: | GLM | Df Residuals: | 6419 |
| Model Family: | Poisson | Df Model: | 5 |
| Link Function: | log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -13579. |
| Date: | Thu, 31 Mar 2022 | Deviance: | 3900.0 |
| Time: | 16:19:46 | Pearson chi2: | 3.48e+03 |
| No. Iterations: | 4 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.7875 | 0.047 | 37.969 | 0.000 | 1.695 | 1.880 |
| attr_me_partners | 0.0146 | 0.004 | 3.445 | 0.001 | 0.006 | 0.023 |
| sinc_me_partners | -0.0057 | 0.004 | -1.542 | 0.123 | -0.013 | 0.002 |
| fun_me_partners | 0.0189 | 0.004 | 5.027 | 0.000 | 0.012 | 0.026 |
| intel_me_partners | -0.0200 | 0.005 | -3.798 | 0.000 | -0.030 | -0.010 |
| amb_me_partners | -0.0011 | 0.003 | -0.354 | 0.724 | -0.007 | 0.005 |

Zaujímavý fakt z tohto modelu je, že čím je človek inteligentnejší, tým menej zaujíma druhého človeka. Zároveň vidíme, že bez akýchkoľvek vlastností, by sme zaujali druhú osobu na hodnotu 1.7 na škále 1-10. Vidíme, že nie každý atribút sa ukázal signifikantný v regresií, ale aj keď sme znovu nafitovali regresiu bez týchto atribútov sa úspešnosť modelu nezväčšila. Rozhodli sme sa preto, pre prehľadnosť dokumentu sem ďalšie tabuľky neuvádzať.

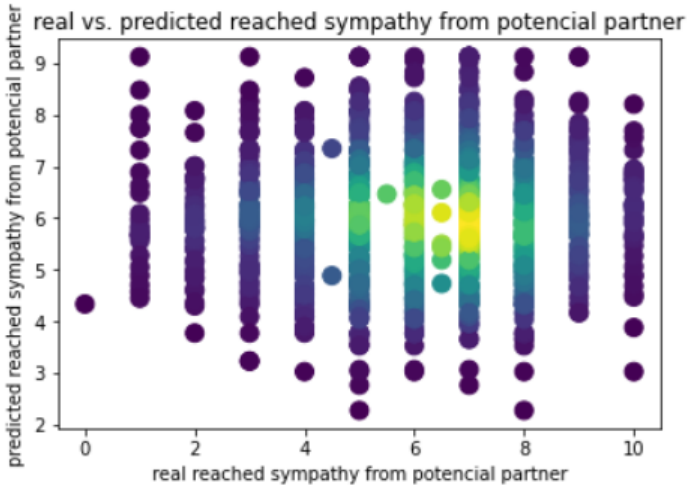
## Analýza parametrov modelu

Ako sme už uviedli, podľa Z štastiky je najsilnejším parametrom pri predikcií ako veľmi sa zapáčime druhému človeka náš zmysel pre humor. Koeficient pri tejto hodnote je 0.0189 - čo značí, že keď sa staneme vtipnejším o jednu jednotku na škále od 1-10, tak sa druhému človeku zapačime o 0.02 viac na škále 1-10. Dané tvrdenie tvrdíme so štandardnou chybou 5%. Vzhľadom k tomu, že konštanta má najväčší vplyv na sympatie je zrejmé, že týchto 5

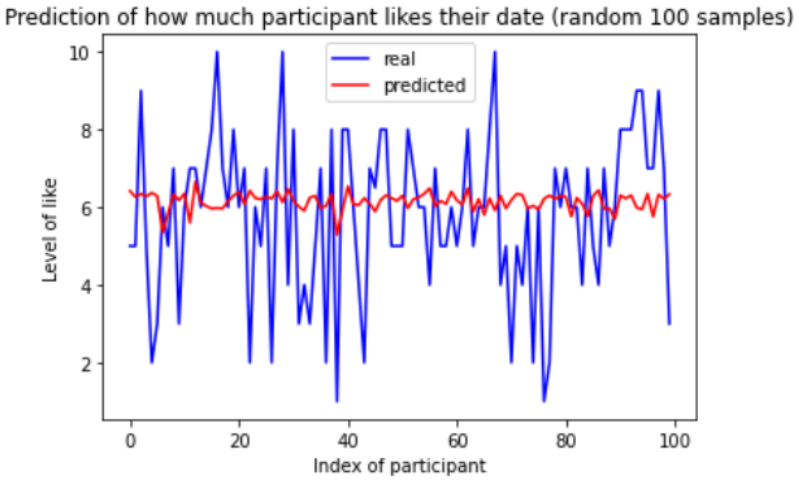
atribútov nie je vhodných pre modelovacie závislosti sympatií a teda, že tento problém je s danými dátami skoro nemožné riešiť.

## Validácia modelu

Vidíme, že závislosť osobnostných atribútov a sympatií druhého človeka existuje, tak sme vykonali na testovacích dátach aj testovaciu inferenciu. RMSE pre našu regresiu bola dosiahnutá na úrovni 1.86 Čo znamená, že pri predikcií sympatií, ktoré získame u druhého človeka sa môže líšiť skoro o 2 jednotky do jednej aj druhej strany. Pri našej škále 1-10 je to spolu rozptyl okolo 4, čiže sa vieme mýliť v rozsahu 40%. Táto hodnota nie je vôbec malá.
Nasledujúci graf zobrazuje vzťah medzi mierou skutočnej a predikovanej dosiahnutej sympatie u opačného pohlavia. Síce naše predikcie majú pomerne veľkú chybovosť, bledšie farby na úrovni stredu osi y hovoria o tom, že náš model zväčša predikoval hodnotu v okolí priemeru možných dosiahnuteľných sympatií.



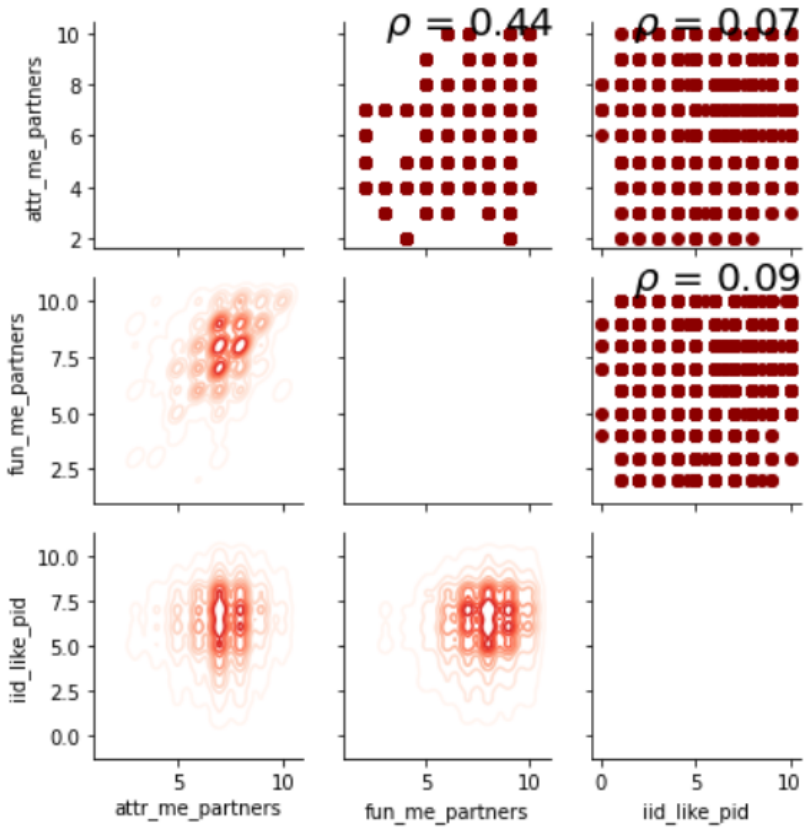

O tom, že náš model nepredikuje mieru dosiahnutej sympatie hovorí aj nasledujúci graf, ktorý prezentuje vzťah medzi mierou skutočnej a predikovanej dosiahnutej sympatie u opačného pohlavia na vybranej vzorke náhodných 100 predikcií.




Vidíme, že náš model založený na Poissonovej regresii predikuje hodnoty mierne okolo priemeru, avšak skutočná variabilita dát je omnoho väčšia.

## Závislosti v prediktoroch

Videli sme, že dva atribúty majú vplyv na sympatie u druhej osoby. V nasledujúcom grafe si môžeme pozrieť závislosti medzi nimi, tretí stĺpec grafu vyjadruje mieru dosiahnutých sympatií u opačného pohlavia.




Medzi našimi nezávislými premennými sú korelácie, vidíme, že zábavnosť sa zvyšuje úmerne s atraktivitou osoby. Očakávateľne, vyššiu koreláciu so sympatiami druhej osoby má atribút hovoriaci o zmysle pre humor, ale stále je to veľmi slabá závislosť.

## Krížová validácia

Výsledný model sme validovali krížovou validáciou, takým spôsobom, že sme si vytvorili 100 krát model natrénovaný na náhodnej 80% množine dát a zvalidovaný na zvyšných 20%. Výsledných 100 hodnôt sme spriemerovali a vypočítali priemernú chybu 1.83 metrikou RMSE. Pričom sme zaregistrovali v dátach štandardnú odchylku 0.03. Priemerná hodnota chyby 1.83 sa môže viazať do oboch strán, takže keď si zoberieme našu škálu 1-10 tak sa môžeme mýliť v 36% celej škály, čo je dosť veľké číslo. Zároveň ako sme spomenuli v predošlej hypotéze, naše dáta nie sú veľmi reprezentatívne, z dôvodu kultúrnych rozdielov a zároveň hodnoty osobných atribútov si ľudia uvádzali sami, čo môže spôsobiť veľkú nepresnosť kvôli subjektívnosti, pokryvenému sebaobrazu. Náš model by niečo dokázal predikovať aj na nových dátach, ale neočakávame výrazne lepšie výsledky. Vďaka tomu, že náš model bol trénovaný na takto zašumených dátach, aj v prípade keby bol testovaný na zašumených dátach, reagoval by rovnako. Teda problém veľkej variance v našom modeli nevidíme. Zároveň sme si vedomí, že náš model má veľký bias, kvôli nejasnej doméne a

dátam. V tomto našom probléme by nám nepomohol ani zjednodušiť model, naopak by sme potrebovali iné črty na modelovanie, nakoľko sa ukázali iba 2 ako signifikantné.

### Zhodnotenie

Zistili sme, že osobnostné vlastnosti jedinca majú vplyv na mieru vyvolaných sympatií u opačného pohlavia, avšak nie je možné presne regresovať túto mieru, nakoľko sa jedná o nie čisto racionálny problém. Z osobnostných atribútov má na mieru získaných sympatií najväčší vplyv práve zmysel pre humor jedinca. Tieto zistenia sú z nášho pohľadu logické, ako sa dá za 4 minúty niekoho zaujať? No predsa tým, že naladíme dobrú atmosféru a rozosmejeme ho.

## Hypotéza 3

*Na základe vlastností človeka je možné určiť jeho preferenciu na inteligenčnú úroveň jeho partnera (akú inteligenciu majú partneri, ktorých si vybrali).*

Predpokladáme, že na základe osobnostných vlastností jednotlivca je možné určiť jeho požiadavky na inteligenciu jeho potenciálneho partnera. Túto hypotézu sme sa snažili overiť rovnakým postupom ako v prípade tretej hypotézy, pomocou Poissonovej a Negatívnej binomiálnej regresie. Zistili sme, že najväčší vplyv na preferencie inteligenčnej úrovne pri našom modelovaní má práve biasová konštanta, čo znamená, že osobnostné atribúty nemajú signifikantnú závislosť s inteligenciou potencionálneho partnera. Sumár regresie môžeme vidieť na nasledujúcom obrázku.

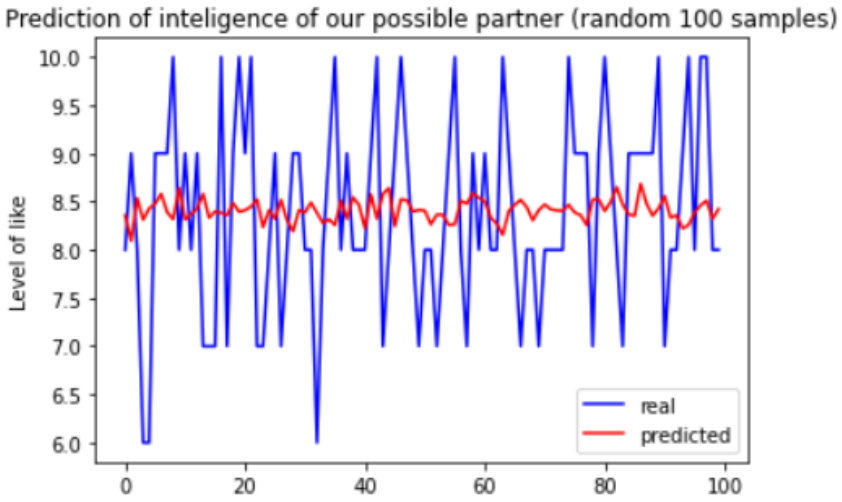
Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | intel_me_partners | No. Observations: | 1052 |
| Model: | GLM | Df Residuals: | 1047 |
| Model Family: | NegativeBinomial | Df Model: | 4 |
| Link Function: | log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -3352.5 |
| Date: | Fri, 01 Apr 2022 | Deviance: | 14.441 |
| Time: | 09:43:19 | Pearson chi2: | 13.6 |
| No. Iterations: | 4 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 2.2063 | 0.306 | 7.206 | 0.000 | 1.606 | 2.806 |
| attr_me | 0.0013 | 0.027 | 0.048 | 0.962 | -0.051 | 0.054 |
| intel_me | -0.0089 | 0.037 | -0.240 | 0.810 | -0.082 | 0.064 |
| sinc_me | -0.0056 | 0.024 | -0.232 | 0.816 | -0.053 | 0.042 |
| amb_me | 0.0047 | 0.021 | 0.220 | 0.826 | -0.037 | 0.046 |

Vidíme, že žiadny z nami zvolených atribútov nie je vhodným pre modelovanie závislosti s inteligenciou partnera. Koeficient pri konštante nám hovorí o tom, že základnou hodnotou, ktorú človek považuje od svojho partnera bez akýchkoľvek iných atribútov je 2.20 z desiatich. Aj napriek tomu, že tento model je vcelku nesignifikantný, vykonali sme inferenciu

na testovacej vzorke kde sme sa dopustili priemernej chyby okolo 1 do každej strany. Vizualizáciu môžete vidieť nižšie.




Vidíme, že náš model síce má chybu všeobecne 20 percent na našej škále od 1-10, avšak vždy predikuje hodnoty len okolo priemeru, čo pre naše dáta s veľkou variancou nie je dostačujúce.

Požadovanú inteligenčnú úroveň nemožno presne určiť na základe 5 osobnostných vlastností jednotlivca z našich dát. Bolo by potrebné mať dostupné väčšie množstvo dát a zobrať iné, presnejšie črty, ale zároveň stále platí, že sympatie a láska na prvý pohľad nemá tendenciu byť racionálna, teda ľahko predpovedateľná.

## Hypotéza 4

*Na základe partnerových preferencií na hobby je možné určiť preferencie na hobby účastníka.*

Vo svete je veľmi zaužívané príslovie, že vrana k vrane sadá. Preto sa nám pri pohľade na naše dáta ukázala otázka, či sa pri toľkých údajoch o záujmoch ľudí stáva, že sa primárne dávajú dokopy ľudia s rovnakými záujmami. Napríklad, či športovci chcú randiť len so športovkyňami.

V tejto hypotéze chcem modelovať hobby účastníka s hobbies jeho partnera. Tým, že sme si zobrali iba ľudí, ktorí prejavili vzájomné sympatie a vieme ich záujmy, vieme sa pozerať na dáta takým spôsobom, že napríklad keď väčšina športovcov má partnerov, ktorí chodia radi do divadla a na turistiku, sú to pre nich dôležité záujmy. Vytvorili sme preto model, kde závislou premennou je to, či je daná osoba športovcom a nezávislými premennými sú záujmy druhého človeka. Testovaním tejto hypotézy očakávame, že koeficient sily záujmov týkajúcich sa napríklad športu bude silnejší ako iných záujmov.

Na začiatok sme boli nútení si dáta predpripraviť, nakoľko participanti prieskumu hodnotili svoje záujmy na škále 1-10, ako veľmi sa tomuto hobby venujú. My sme pre naše potreby potrebovali ľudí na tých, ktorí sa tomu záujmu jednoducho venujú a tých, ktorí nie. Zvolili sme preto delenie na základe priemeru a teda sme si vypočítali priemernú hodnotu pre každý atribút naprieč celým datasetom a ak osoba sa hodnotila v danom hobby nadpriemerne, dostala hodnotenie 1 - záujmu sa venuje, v opačnom prípade 0 - záujmu sa

nevenuje. Takto sme dostali predpripravený dataset núl a jednotiek, ktorý hovorí o záujmoch, ktorým sa ľudia venujú. V datasete sa nachádzalo zopár riadkov ktoré mali hodnoty NaN, tie sme z datasetu vylúčili. Nakoľko sme mali hodnoty len 0 a 1, nešetrili sme žiadne vybočujúce hodnoty. Všetky experimenty boli robené len na subsete datasetu, ktorý obsahoval len ľudí, ktorí sa označili, že majú vzájomné sympatie.

V nasledovnom reporte môžete vidieť štatistiky logistickej regresie pre ľudí, ktorí boli označení ako športovci.

```
sports
Optimization terminated successfully.
         Current function value: 0.680532
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                 sports   No. Observations:                 8378
Model:                          Logit   Df Residuals:                     8360
Method:                           MLE   Df Model:                           17
Date:                Thu, 31 Mar 2022   Pseudo R-squ.:                0.01132
Time:                        23:30:37   Log-Likelihood:                -5701.5
converged:                       True   LL-Null:                       -5766.8
Covariance Type:            nonrobust   LLR p-value:                 1.438e-19
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                0.0181      0.070      0.257      0.797      -0.120       0.156
sports_partners     -0.1600      0.051     -3.150      0.002      -0.260      -0.060
tvsports_partners   -0.0612      0.050     -1.216      0.224      -0.160       0.037
exercise_partners    0.0237      0.049      0.489      0.625      -0.071       0.119
dining_partners      0.0425      0.050      0.849      0.396      -0.056       0.141
museums_partners    -0.0888      0.071     -1.256      0.209      -0.227       0.050
art_partners         0.0611      0.069      0.885      0.376      -0.074       0.196
hiking_partners      0.1747      0.049      3.594      0.000       0.079       0.270
gaming_partners     -0.2433      0.048     -5.024      0.000      -0.338      -0.148
clubbing_partners   -0.0119      0.048     -0.251      0.802      -0.105       0.081
reading_partners     0.0960      0.048      1.992      0.046       0.002       0.190
tv_partners          0.1244      0.049      2.523      0.012       0.028       0.221
theater_partners     0.2166      0.055      3.939      0.000       0.109       0.324
movies_partners     -0.0312      0.054     -0.583      0.560      -0.136       0.074
concerts_partners   -0.0460      0.056     -0.818      0.414      -0.156       0.064
music_partners       0.0251      0.052      0.480      0.631      -0.077       0.127
shopping_partners    0.0991      0.050      1.973      0.048       0.001       0.197
yoga_partners        0.0896      0.048      1.884      0.060      -0.004       0.183
==============================================================================
```

Vidíme, že zo všetkých 16 atribútov vyšlo signifikantných podľa P-value len 5 atribútov, pričom podľa Z- štatistiky sú najsignifikantnejšie záujmy ľudí chodiť do divadla a chodiť na turistiku. Koificienty pri týchto atribútoch hovoria, že keď osoba rada chodí do divadla alebo na turistiku, tak je šanca, že sa participanti budú vzájomne páčiť väčšia o zhruba 20 percent. Model môžeme považovať za signifikantný nakoľko P-values niektorých atribútov sú validné a taktiež pseudo R-square hodnota vyjadrujúca pomer, koľko variability v predikovanej premennej náš model nezachytí, sa blíži k nule. Signifikancia konštanty je takmer nulová, takže športovci by chceli chodiť s ľudmi, ktorý nemajú žiadne záujmy s pravdepodobnosťou 2%. Celkové výsledoky sú sčasti prekvapivé, nakoľko sme čakali, že najsilnejší atribút bude práve hobby športu, no turistika sa taktiež klasifikuje do športu. Zaujímavejším pozorovaním je však záujem chodiť do divadla.

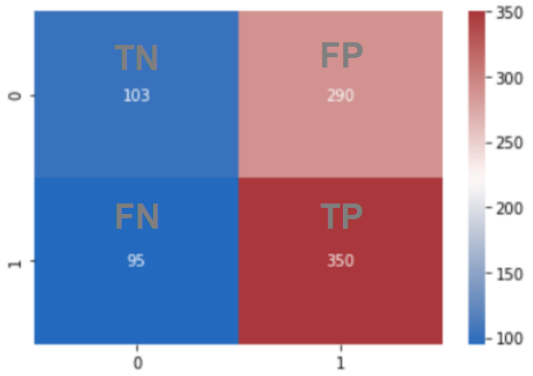
Podobné pozorovanie sme vykonali taktiež napríklad na umelcoch.

```
art
Optimization terminated successfully.
         Current function value: 0.674201
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                    art   No. Observations:                 8378
Model:                          Logit   Df Residuals:                     8360
Method:                           MLE   Df Model:                           17
Date:                Thu, 31 Mar 2022   Pseudo R-squ.:                0.004689
Time:                        23:30:38   Log-Likelihood:                -5648.5
converged:                       True   LL-Null:                       -5675.1
Covariance Type:            nonrobust   LLR p-value:                 1.320e-05
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                  0.3431      0.071      4.834      0.000       0.204       0.482
sports_partners        0.1957      0.051      3.833      0.000       0.096       0.296
tvsports_partners      0.0129      0.051      0.255      0.798      -0.086       0.112
exercise_partners     -0.0518      0.049     -1.060      0.289      -0.147       0.044
dining_partners        0.0534      0.051      1.057      0.290      -0.046       0.152
museums_partners      -0.0766      0.072     -1.070      0.285      -0.217       0.064
art_partners           0.0875      0.070      1.255      0.209      -0.049       0.224
hiking_partners       -0.1039      0.049     -2.122      0.034      -0.200      -0.008
gaming_partners        0.0994      0.049      2.037      0.042       0.004       0.195
clubbing_partners      0.0210      0.048      0.439      0.661      -0.073       0.115
reading_partners       0.0065      0.049      0.134      0.893      -0.089       0.102
tv_partners           -0.0814      0.050     -1.641      0.101      -0.179       0.016
theater_partners      -0.1426      0.056     -2.558      0.011      -0.252      -0.033
movies_partners        0.1204      0.054      2.225      0.026       0.014       0.227
concerts_partners      0.0170      0.057      0.300      0.764      -0.094       0.128
music_partners        -0.0096      0.053     -0.182      0.856      -0.113       0.093
shopping_partners     -0.0836      0.051     -1.654      0.098      -0.183       0.015
yoga_partners         -0.0507      0.048     -1.060      0.289      -0.144       0.043
==============================================================================
```

Vidíme, že sa ukázalo, že umelci sa veľmi často chcú dať do vzťahu so športovcami a s ľuďmi, ktorí chodia do kina. Opäť, pozorovanie ohľadom športovcov nás mierne prekvapilo. Tento model považujeme opäť za signifikantný, kvôli týmto dvom atribútom s priaznivou P-value a pseudo R-square hodnotou takmer 0. Zároveň vidíme, že signifikancia konštanty je v tomto prípade nezanedbateľná. Dokonca je najväčšia zo všetkých atribútov. To znamená, že umelcom menej záleží na tom aby mal druhý človek hoci len nejaké záujmy. To je veľmi zaujímavé, nakoľko umelci sú všeobecne emočne založení, že radi zdieľajú svoje záujmy.

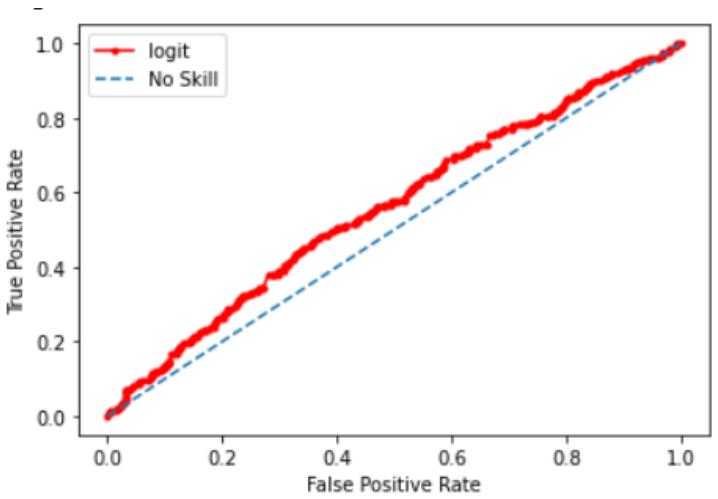
## Validácia

Natrénovali sme náš model na 90% dát a testovali na zvyšku datasetu. Keď sme si zobrali iba predikciu, či je náš nový partner športovec dosiahli sme takéto výsledky.




Accuracy: 54.06%
Recall: 78.65%
Precision: 54.69%
F1_score: 64.52%
Specificity: 26.21%
Sensitivity: 78.65%

Vidíme, že sme dosiahli úspešnosť predikcie 54%, avšak keďže riešime problém s nevybalancovaným datasetom, musíme sa pozerať na iné metriky. Správne sme klasifikovali každých štyroch z piatich športovcov a viac ako polovica ľudí, ktorých sme za športovcov označili, nimi skutočne boli. Vidíme, že hodnota specificity, ktorá hovorí o správnej klasifikácií nešportovcov je veľmi malá, to znamená, že uprednostňujeme klasifikáciu športovcov. Aj keď v tomto prípade, je pre nás hodnota falošne pozitívnych a falošne negatívnych rovnaká, stále sme spokojní s dosiahnutými výsledkami, nakoľko keby optimalizujeme podľa F1 skóre, znížili by sme pokrytie športovcov a to by sme nechceli. Taktiež sme vytvorili ROC krivky.




Očakávateľne sa naša ROC krivka nevzďaľuje od diagonálnej krivky, čo znamená, že stále dosahujeme len o niečo lepšie výsledky ako náhoda. Celkovo sme dosiahli hodnotu ROC AUC skóre málo nad hodnotou 56%.

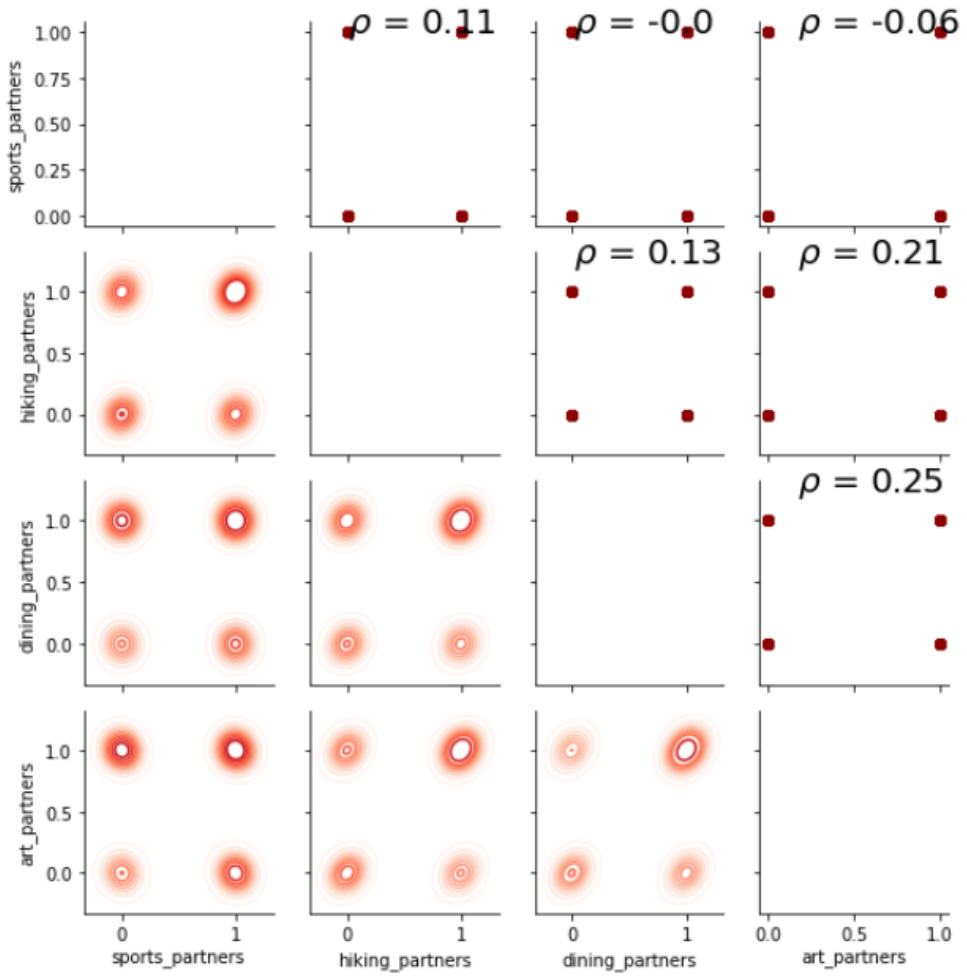
## Krížová validácia

Predošlé prezentované výsledky sme ešte potrebovali overiť dôkladnejším validovaním, tak sme použili krížovú validáciu ako pri ostatných hypotézach v tomto projekte. Vykonali sme teda 100 náhodných tréningov. Dosiahli sme priemernú úspešnosť 56% pre ROC AUC skóre, čo nie je až také zlé na naše predchádzajúce výsledky. Táto hodnota bola nameraná s 1% štandardnou chybou. Rovnako, ako sme spomínali v predchádzajúcich hypotézach, dáta nie sú z uvedených dôvodov reprezentatívne. Takisto vzhľadom k tomu, že riešime úzko špecifikovaný problém, tento model by nebol vhodný pre iný dataset. Použili sme

zároveň asi najjednoduchší model, aký vieme použiť. Komplexnejšie modely ako napríklad neurónové siete, by mohli byť pri tomto priamočiarom probléme náchylné na pretrénovanie.

## Kolinearita dát

Musíme povedať, že očakávateľne sa v našich dátach nachádzajú kolinearity. Je to veľmi logické nakoľko už len ľudia, čo radi športujú, budú aj radi chodiť na turistiky.




V tomto grafe vidíme zopár kolinearít v našich dátach. Vidíme, že keď napríklad ľudia radi chodia na večere do reštaurácií, tak majú aj radi umenie. Asi chodia po umeleckých vystúpeniach na romantické večere.

## Zhodnotenie

Na záver môžeme len konštatovať, že počiatočné tvrdenie vrana k vrane sadá sa nám nepotvrdilo, no skôr by sme dovolili tvrdiť, že na základe našich dát sa protiklady priťahujú. Takýto výsledok nám však príde tiež logický, pretože aspoň sa vedia páry pri odlišných záujmoch obohatiť o nové zážitky.

# Hypotéza 5

*Technickí inžinieri inklinujú k medičkám*

Podobným spôsobom ako sme sa pozerali na predošlú hypotézu, sme skúsili overiť hypotézy ohľadom špecifických profesií, či sa viac priťahujú. Primárne sme chceli overiť tvrdenie, že technickí inžinieri inklinujú k medičkám. Vykonali sme preto one-hot encoding nad kariérnym zameraním. Museli sme ale použiť zovšeobecnenú regresiu, lebo logistická nebola schopná v tak riedkych dátach skonvergovať k výsledku, čo vyústilo k slabšiemu modelu a aj všetky atribúty mali záporné koeficienty. Detaily tohto experimentu možno vidieť v notebooku, rozhodli sme sa tu prezentovať jeden sumár z regresie pre doktorov.

doctor_me

Generalized Linear Model Regression Results

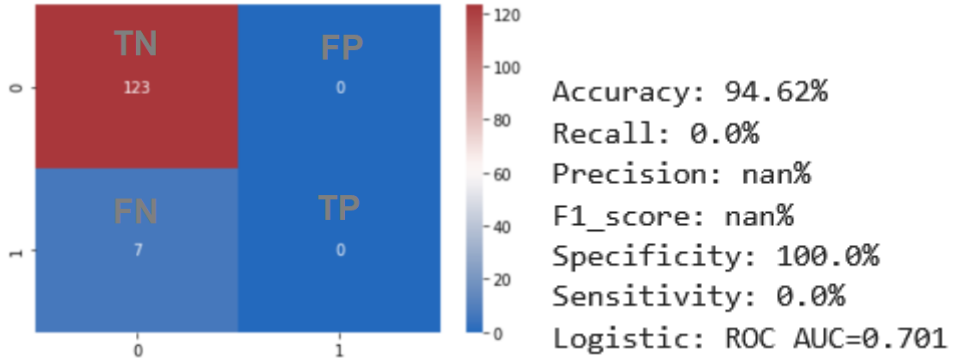
| | | | |
|---|---|---|---|
| Dep. Variable: | doctor_me | No. Observations: | 1296 |
| Model: | GLM | Df Residuals: | 1280 |
| Model Family: | NegativeBinomial | Df Model: | 15 |
| Link Function: | log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -268.39 |
| Date: | Fri, 01 Apr 2022 | Deviance: | 345.47 |
| Time: | 00:14:13 | Pearson chi2: | 1.15e+03 |
| No. Iterations: | 19 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| lawyer | -2.9178 | 0.419 | -6.961 | 0.000 | -3.739 | -2.096 |
| academic | -2.8416 | 0.224 | -12.658 | 0.000 | -3.282 | -2.402 |
| psychologist | -3.1355 | 0.722 | -4.341 | 0.000 | -4.551 | -1.720 |
| doctor | -4.2195 | 1.007 | -4.189 | 0.000 | -6.194 | -2.245 |
| engineer | -3.5264 | 1.015 | -3.476 | 0.001 | -5.515 | -1.538 |
| entertainment | -2.9124 | 0.459 | -6.342 | 0.000 | -3.812 | -2.012 |
| bussiness | -3.1987 | 0.273 | -11.731 | 0.000 | -3.733 | -2.664 |
| real estate | -22.6261 | 2.03e+04 | -0.001 | 0.999 | -3.98e+04 | 3.97e+04 |
| international | -2.4953 | 0.368 | -6.783 | 0.000 | -3.216 | -1.774 |
| undecided | -2.5177 | 0.465 | -5.416 | 0.000 | -3.429 | -1.607 |
| social work | -2.7300 | 0.596 | -4.582 | 0.000 | -3.898 | -1.562 |
| speach | -22.6261 | 4.97e+04 | -0.000 | 1.000 | -9.74e+04 | 9.73e+04 |
| politics | -1.7047 | 0.769 | -2.218 | 0.027 | -3.211 | -0.198 |
| athletics | -22.6261 | 4.97e+04 | -0.000 | 1.000 | -9.74e+04 | 9.73e+04 |
| other | -2.6391 | 1.035 | -2.550 | 0.011 | -4.668 | -0.610 |
| journalism | -22.6261 | 2.48e+04 | -0.001 | 0.999 | -4.87e+04 | 4.86e+04 |
| architecture | 0 | 0 | nan | nan | 0 | 0 |

Vidíme, že všetky atribúty majú negatívny koeficient, čo znamená, že doktori nemajú záujem o vzťah s nejakým povolaním výhradne. Toto môže byť spôsobené veľmi riedkymi dátami, vďaka čomu regresia nefunguje optimálne. Zároveň z EDY vieme, že máme veľmi veľa druhov povolaní s málo početnosťami. Preto túto hypotézu nevieme dostatočne overiť.

## Validácia

Skúsili sme taktiež tento model zvalidovať na testovacích dátach. Bohužial, vidíme z predchádzajúceho fitu, že sa nevie naučiť naše dáta. Tento náš model predikuje každému particitantovi, že nemá záujem o vzťah so žiadnym povolaním. Uvázdame confusion matrix.




Vidíme, že všetky hodnoty sa predikujú na nulu a tento model je teda absolútne nepoužiteľný. Túto hypotézu teda na základe dát nevieme overiť.

# Zhodnotenie

Pokúsili sme aplikovať metódy dátovej analýzy na dáta z experimentu speed datingu. Overili sme niektoré hypotézy, ktoré sa nám zdali ako logické na overenie z pohľadu na dáta, no môžeme povedať, že sme očakávali trochu lepšie výsledky. Na druhej strane vieme, že predikovať veličiny v doméne randenia a lásky je takmer nemožné z dôvodu veľkej iracionality v sympatiách medzi ľuďmi. V modelovaní sme dosiahli čiastočne očakávané hodnoty, ktoré podporovali našu intuíciu, ale nie v takej miere, aby sme mohli povedať, že na základe týchto modelov vieme predikovať potenciálnu náklonnosť medzi jednotlivcami.