Zadanie 1

Spracovanie a príprava dát

Vymazal som stĺpce: D_appid, D_name, pretože hodnota pre každý riadok je jedinečná a tak to nemá vplyv na to ci je hra zadarmo alebo nie. Ďalej som vymazal stĺpce Well-Written, Masterpiece, Lore-Rich, Epic, Emotional, Cult Classic, Competitive, Beautiful, english, pretože v týchto stĺpcoch prevažuje iba jedna hodnota,

Zo stĺpca D_reviews na číselne hodnoty, aby som ich mohol dať do trénovania. Číslo 0 som priradil najhoršiemu hodnoteniu a najlepšiemu číslo 8.

Vymazal som riadky s nulovými hodnotami a duplicitné riadky.

Zo stĺpca D_release_date som vybral rok vydania hry, ktorý som uložil do nového stĺpca s názvom year, následne som D_release_date vymazal.

Vypočítal som priemer pre každý interval zo stĺpca D_owners a vytvoril nový stĺpec owners, ktorý reprezentuje počet majiteľov hry. Stĺpec D_owners som potom vymazal.

Z D_tags som hľadal tagy, ktoré by boli dostatočne zastúpené v hrách. A rozhodol som sa, že vyberiem tag, ktorý označuje, či je hra indie. Vytvoril som stĺpec indie do ktorého som dal hodnoty 0 alebo 1 podľa toho, či je hra indie alebo nie.

Zo stĺpca D_genre som dostal žáner hry. Vytvoril som nový stĺpec genres kde som uložil žánre jednotlivých hier.

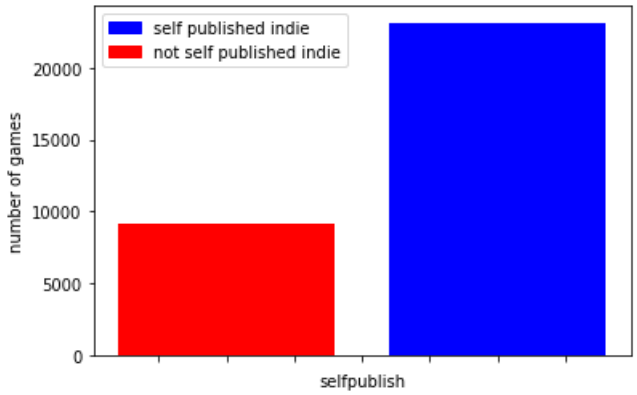
Pomocou Label encoderu som previedol stĺpce D_developer, D_publisher a genres na číselné hodnoty.

Normalizoval som dáta pre stĺpec ccu a owners pomocou funkcie reshape.

Aby som mal vyrovnané dáta tak som vymazal 37150 riadkov s platenými hrami.

EDA

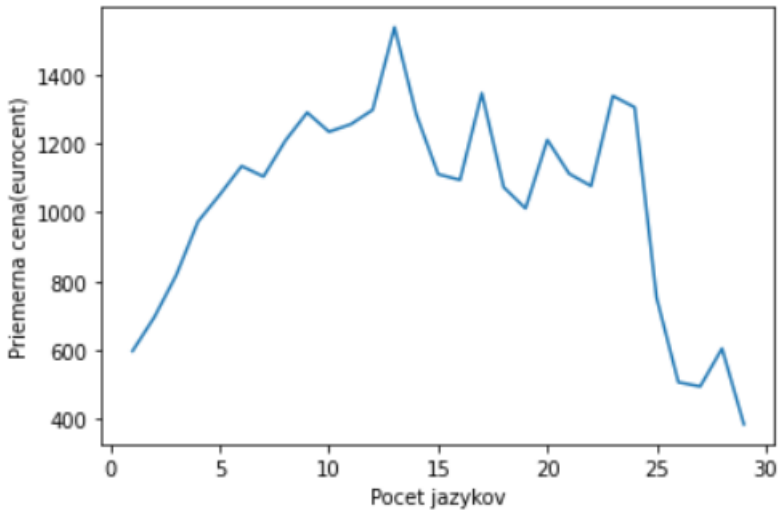
Aký je súvis medzi self-published a indie?



Na obrázku môžeme vidieť 2 stĺpce, ktoré predstavujú počet indie hier. Červený stĺpec označuje, že hra nie je self-published, modrý stĺpec označuje, že hra je self-publish aj indie. Not self published indie hier je skoro 10000, zatiaľ čo self-published indie hier je skoro 25 000. Z grafu vieme vyčítať, že

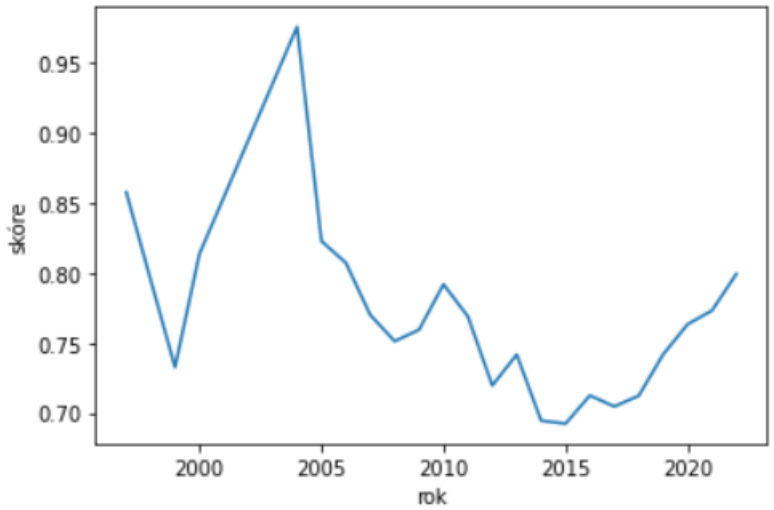
väčšina indie hier je self-published. Je viac ako dvojnásobná pravdepodobnosť, že ak je hra indie tak bude aj self-published.

Aká je závislosť medzi priemernou cenou a počtom jazykov?



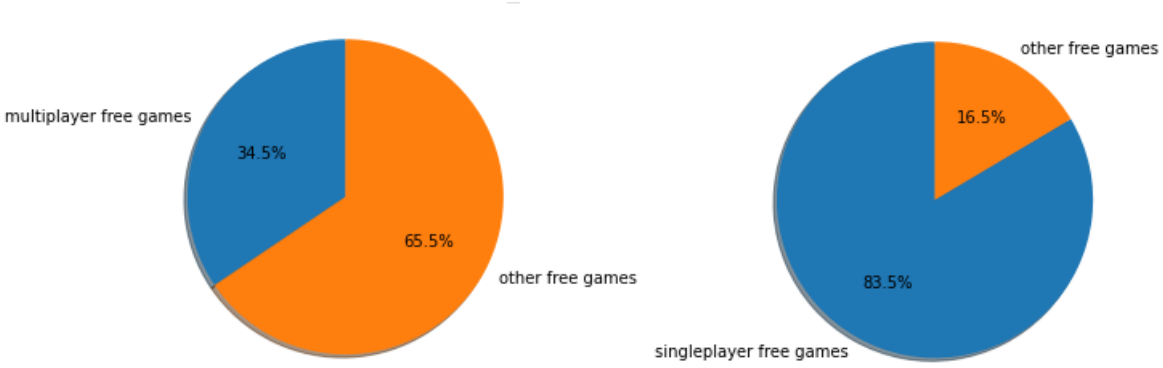
Na grafu je znázornená ako sa vyvíja cena s pribúdajúcim počtom jazykov do ktorých bola hra preložená. Hry som si rozdelil do skupín podľa toho do koľkých jazykov boli preložené a následne som si vypočítal priemerné ceny týchto hier. Na grafe môžeme vidieť, že cena hry rastie do 14 jazykov, kde je priemerná cena najvyššia. Ak je hra preložená do 25 jazykov a viac tak cena rapídne klesá.

Ako sa vyvíjalo skóre hier v priebehu rokov?



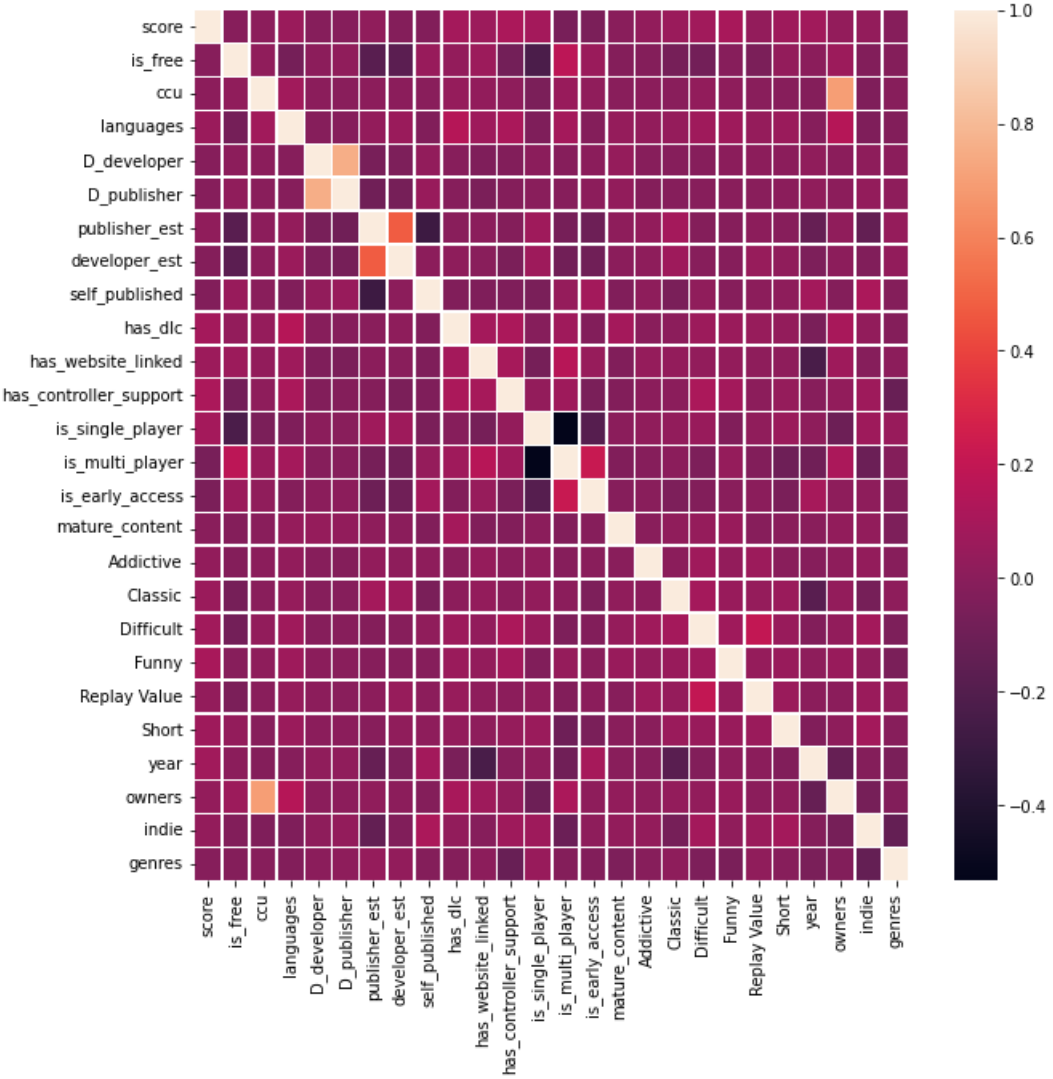
Na obrázku môžeme vidieť aké priemerné skóre mali hry podľa rokov. Najlepší pomer medzi kladnými hodnoteniami a zápornými hodnoteniami mali hry pred rokom 2005. Od tohto roku skóre klesalo až do roku 2015 kedy dosiahlo minimum. Následne po roku 2015 sa skóre hier zvyšuje.

Je súvis medzi tým ak je hra singleplayer a zadarmo?



Na prvom obrázku máme znázornené všetky hry, ktoré sú zadarmo. Z nich tvorí 34,5% multiplayer hry. Na druhom obrázku máme znázornené akú veľkú časť predstavujú singleplayer hry z hier ktoré sú zadarmo. Je ich 83,55% čo znamená, že veľkú časť hier zadarmo tvoria hry, ktoré sú aj singleplayer.

Korelačná matica

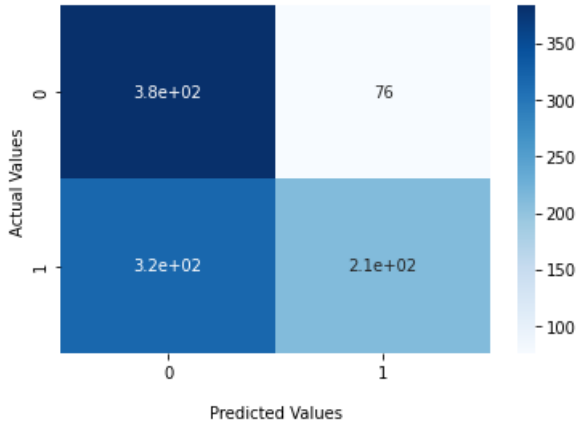


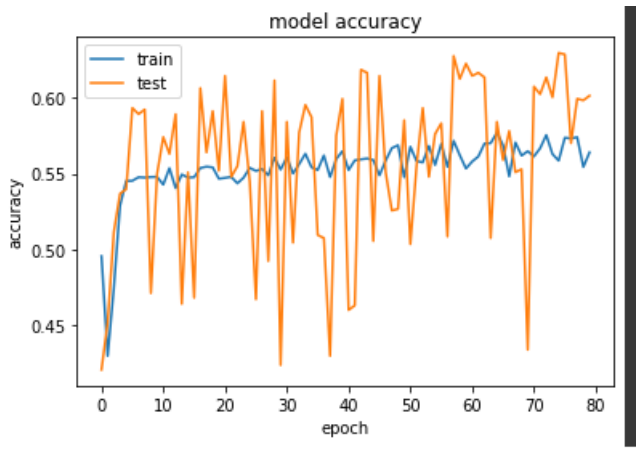
Na obrázku môžeme vidieť graficky znázornenú korelačnú maticu. Vyjadruje závislosti jednotlivých stĺpcov od druhých stĺpcov. Napríklad vidíme silnú závislosť medzi stĺpcami ccu a owners. Taktiež zápornú závislosť medzi tým či je hra singleplayer alebo multiplayer. Ďalšia veľká závislosť je medzi self_published a published_est.

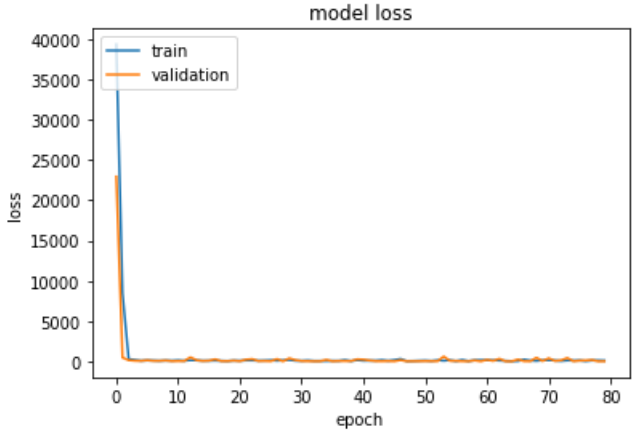
Trénovanie neurónovej siete

Po spracovaní a úprave dát môžeme začať trénovať neurónovú sieť. Na tvorbu neurónovej siete som použil knižnicu Keras. Vytvoril som sieť s 25 vstupmi. Pridal som 2 skryté vrstvy s 22 a 15 neurónmi. Na prvú skrytú vrstvu som dal aktivačnú funkciu relu. A na výstupnú vrstvu som dal 1 neurón s aktivačnou funkciou sigmoid. Optimalizátor som si zvolil adam s learning rate 0,0001. Ako chybovú funkciu som použil binary_crossentropy, pretože dáta delím do 2 tried. Z testovacích dát som vyčlenil 10% pre validačné dáta. Počet epoch som nastavil na 150 a early stoping na 15 epoch.

Trénovanie prebiehalo nasledovne:



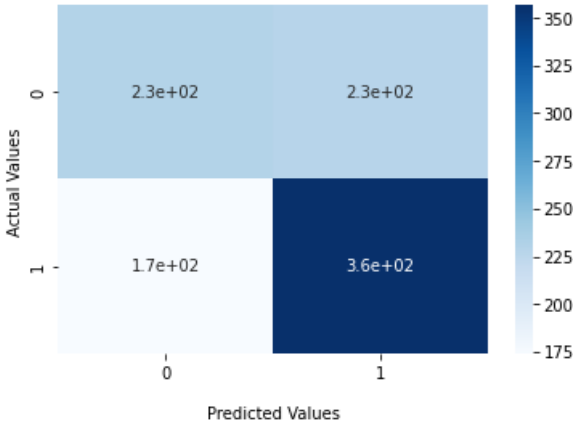


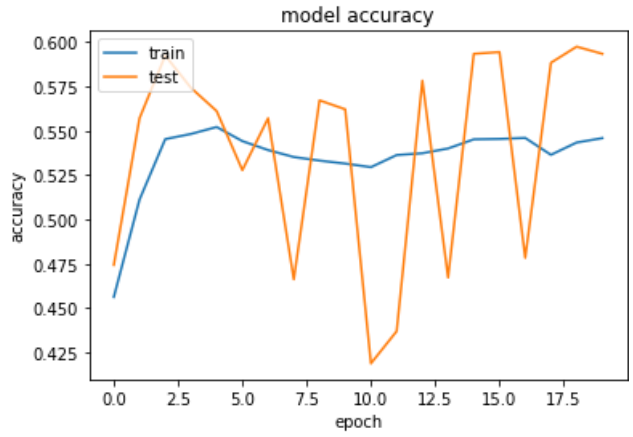


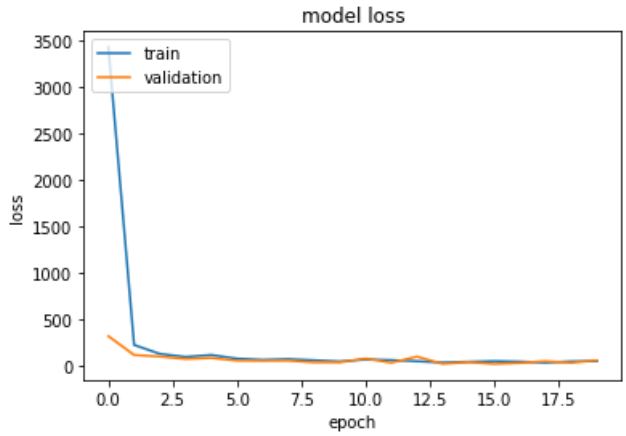
Vidíme, že krivka trénovania je zo začiatku veľmi strmá a následne sa drží v jednej rovine, čo nie je dobré trénovanie. Z konfúznej matice vidíme, že sieť vie na validačných dátach dobre roztriediť hry, ktoré sú platené. Celková presnosť neurónovej siete je 60%.

Druhé trénovanie:

Zmenil som počet neurónov v skrytých vrstvách na 18 a 13 a pridal som do druhej vrstvy aktivačnú funkciu relu.

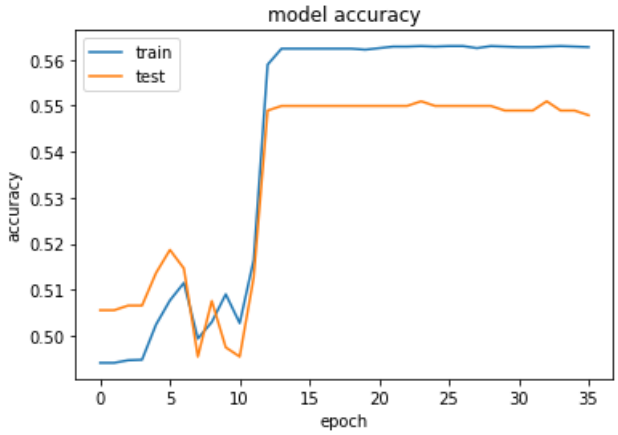


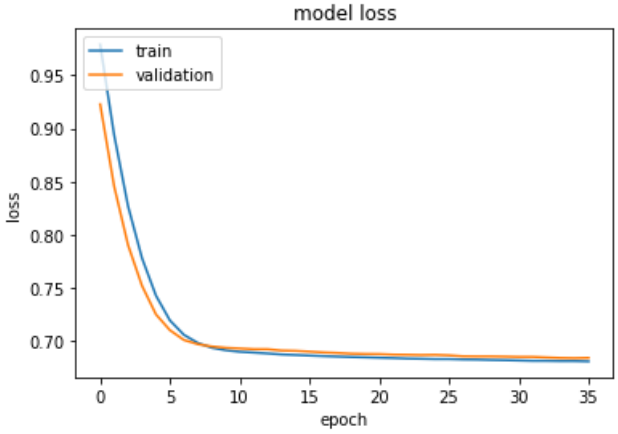




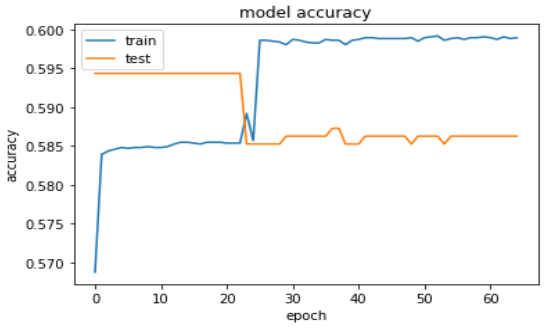
Trénovanie sa zastavilo hneď po 20 epochách. Krivka je stále strmá, ale sieť sa zlepšila v určovaní hier, ktoré sú zadarmo. Celková úspešnosť 59%.

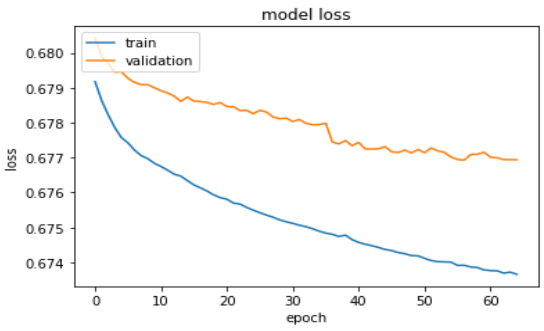
Ďalšie trénovanie: vytvoril som modelcheckpoint kde si ukladám najlepší model neurónovej siete. Následne tento model checkpoint využívam v callbackoch. Skúsil som meniť lerning_rate na 0,00001. Nezaznamenal som nejaký lepší výsledok aj keď krivka sa zlepšila a klesá postupne k nižšej chybovosti.



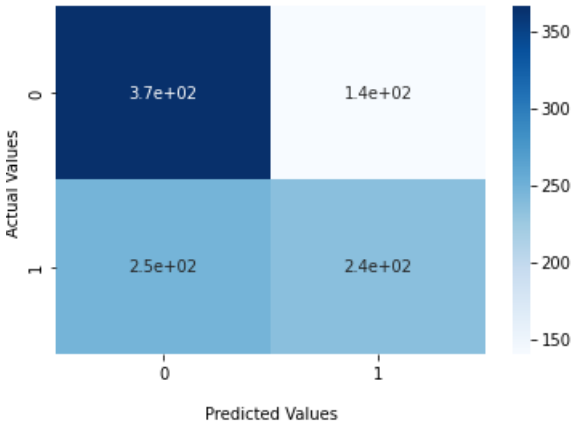


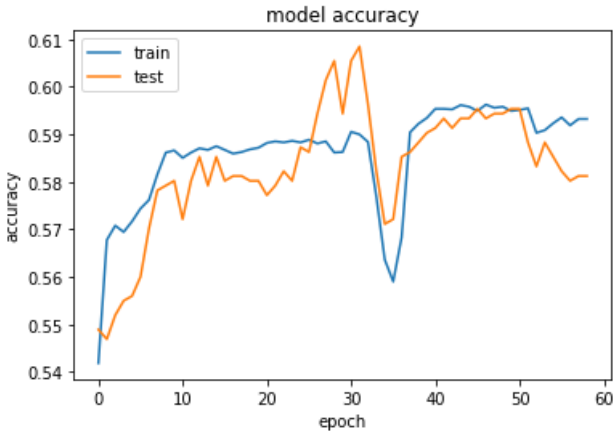
Ďalšie trénovanie: Zmenil som aktivačnú funkciu v prvej skrytej vrstve na sigmoid.

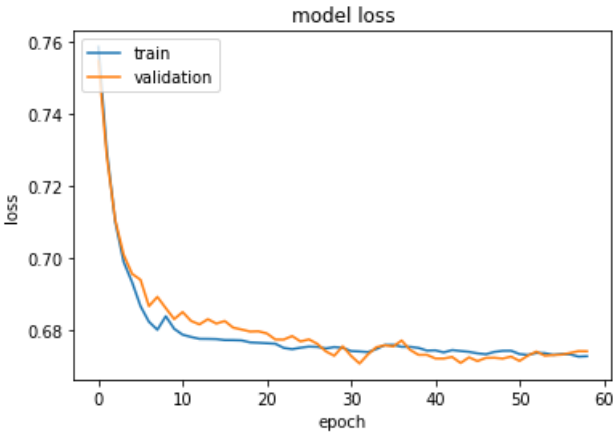




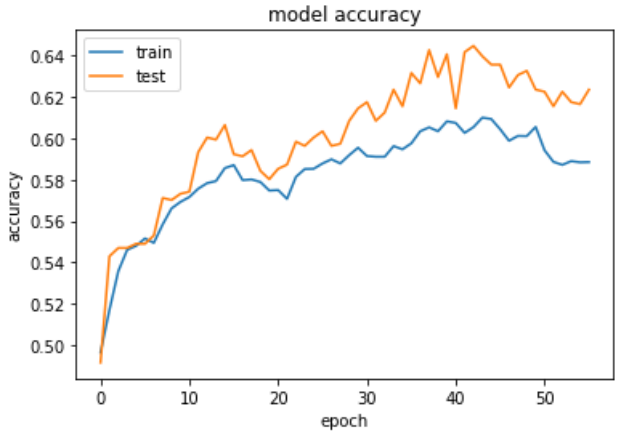
Ďalšie trénovanie: Zmenil som learning rate späť na 0,001. Zmenil som aktivacne funkcie na relu, sigmoid a sigmoid. Úspešnosť 60,8%.

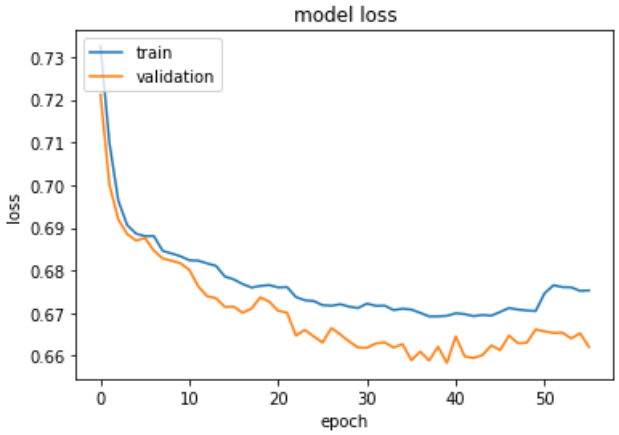


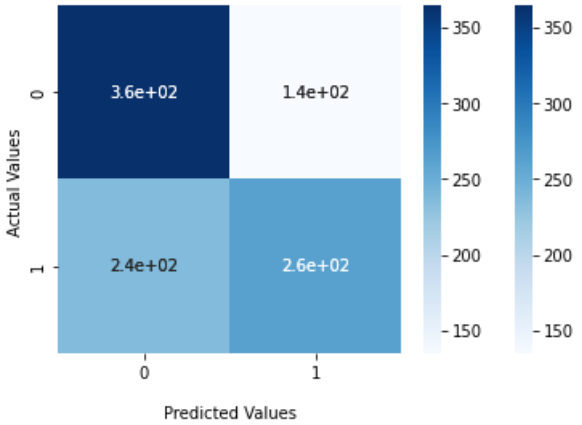




Ďalšie trénovanie: Zmenil som počet neurónv v skrytých vrstvách na 18 a 14. Celková úspešnosť 64,5%.


model accuracy
train
test
accuracy
epoch

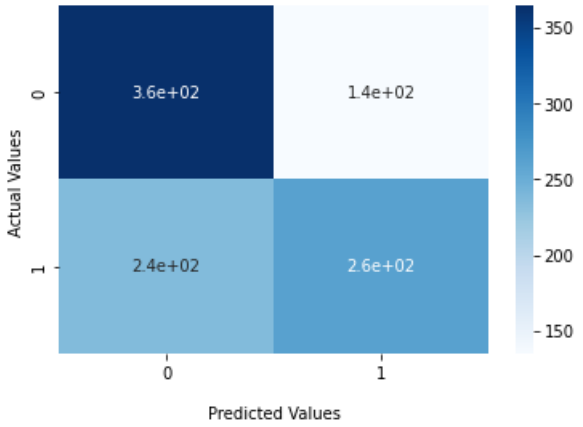

model loss
train
validation
loss
epoch


3.6e+02
1.4e+02
2.4e+02
2.6e+02
Actual Values
Predicted Values

Testovacie dáta – 3 skryté vrstvy . prvá 19 neutónov, druhá 15. Aktivačné funkcie sigmoid.

Úspešnosť 62,8%

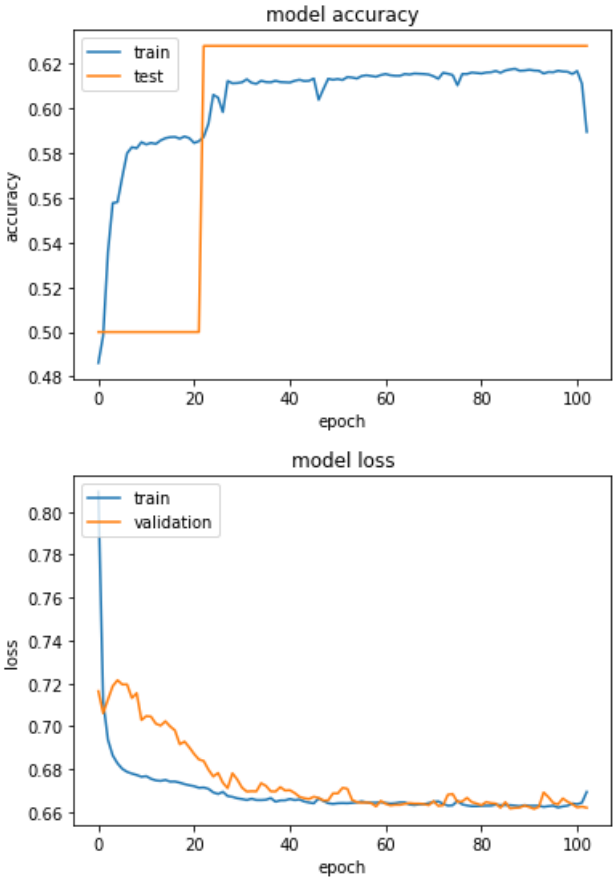
Konfúzna matica pre výsledky testovania.



Trénovacie dáta

3042 1454]

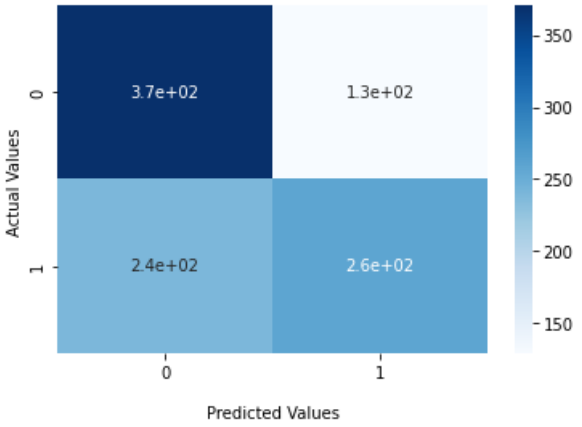
2166 2257



Testovacie data – 3 skryté vrstvy. V prvej 18 neurónov, druhá vrstva 13 neurónov, tretia vrstva 9 neurónov. Prvá vrstva aktivačná funkcia sigmoid, posledná vrstva aktivačná funkcia sigmoid

Úspešnosť 63%.

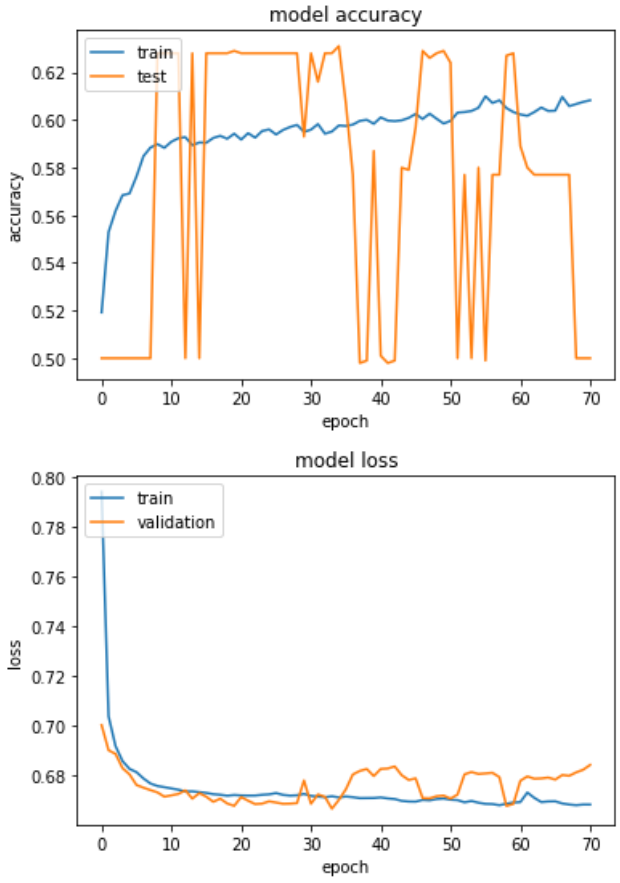
Konfúzna matica pre testovacie dáta



Konfúzna matica pre trénovanie dáta.

2490 2006

1593 2830

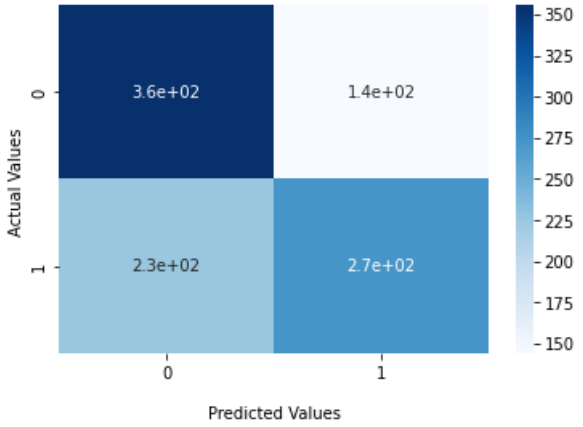


Testovacie data

4 skryté vrstvy. Prvá vrstva 18 neurónov aktivačná funkcia sigmoid, druhá vrstva 13 neurónov, aktivačná funkcia relu, tretia vrstva 9 neurónov, posledná vrstva aktivačná funkcia sigmoid.

Úspešnosť 63%.

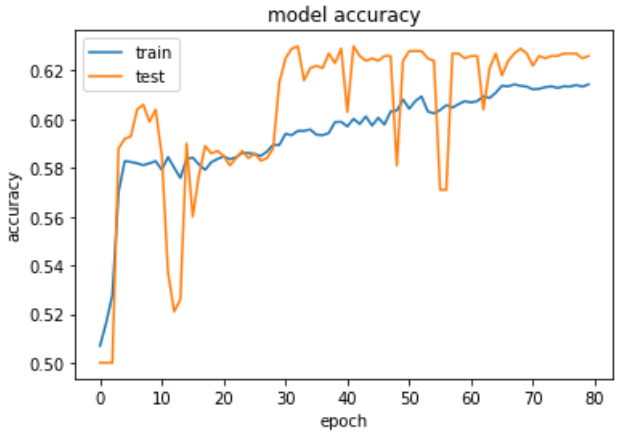
Testovacie data

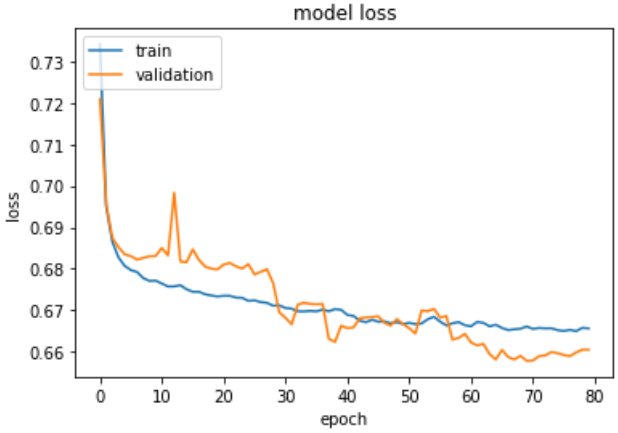


Trénovacie data

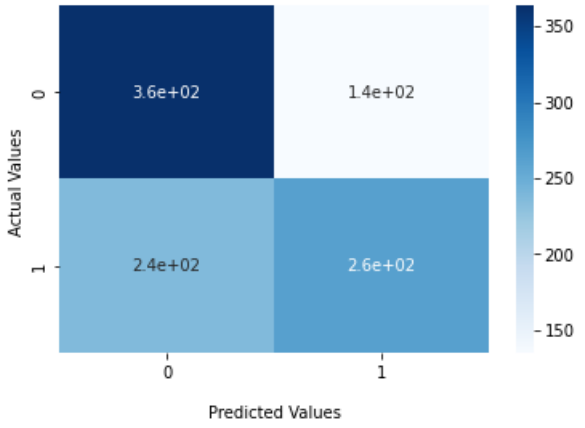
2628 1868

1737 2686



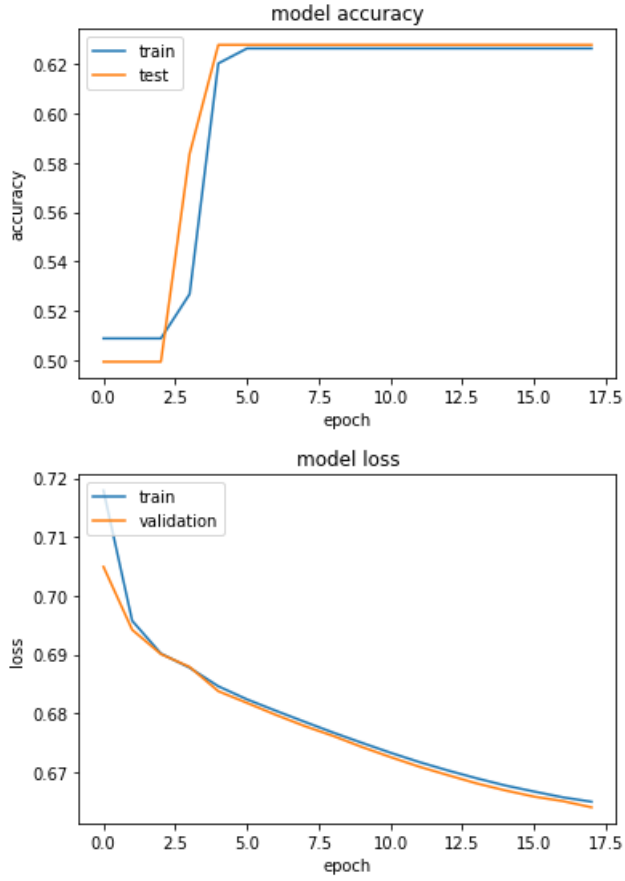


Priebeh trénovania po vymazaní stĺpcov publisher a developer. 62,7%



[[3262 1277]

[2057 2323]]



Ďalšie trénovania

Zmena Learning ratu mala vplyv na to aká strmá bola krivka učenia, pri veľmi malej hodnote krivka klesala pomaly a často sa stalo, že sa hodnota ani nemenila. Naopak pri väčšom čísle, graf neklesal rovnomerne ale kolísal.

Pri malom počte epoch, došlo k tomu, že sa sieť nestihla dostatočne natrénovať. Pri použití iba relu funkcii, sa mala sieť tendenciu zaseknúť iba v jednom minime a ďalej sa nezlepšovať. Pri príliš veľkom počte neurónov sa sieť naučila vynikajúco roztrieďovať iba jednu skupinu hier a úspešnosť bola okolo 50%. Bez checkpointov brala sieť len posledný výsledok, ktorý nebol vždy najlepším. Ak som odstránil early stoping tak sa sieť pretrénovala.