I-SUNS: Zadanie č. 1

Neurónové siete

Steam hry

Matúš Vetrík

Úlohou práce je natrenóvať neurónovú sieť, ktorá má na základe dát z hernej platformy Steam zistiť, či sa jedná o hru zadarmo alebo platenú.

Na riešenie tohto zadania bol zvolený programovací jazyk Python a viacero pomocných knižníc. Na Na úpravu a spracovanie dát boli použité knižnice **Pandas** a **Numpy**. Na graficke zobrazenie dát boli použité knižnice **Plotly** a **Matplotlib.** Na trénovanie neurónovej siete bola použitá knižnica **Tensorflow**. Projekt je vytváraný v prostredí **Jupyter**.

Spracovanie dát

Steam dáta nie sú vo forme vhodnej na trénovanie. Bolo potrebné urobiť viacero úprav dát aby bol dataframe v správnej forme.

Ako prvé sme naformátovali dáta, ktoré boli v tabulke v zlom formáte. Stĺpec "D_release_date" predstavuje dátum vydania danej hry. Obsahoval dátum v tvare "den mesiac, rok". Z tohto stĺpca sme vytiahli iba hodnotu rok a pretypovali sme ju zo stringu na integer. Ďalšou úpravou prebehol stĺpec "D_owners". V tomto stĺpci sú udávané intervali koľko užívateľov vlastní hru. Hodnota je v tvare "x .. y". Z intervalu vytathujeme vyššiu hodnotu intervalu, ktorá je pretypovaná zo stringu na integer. Následne upravujeme hodnoty stĺpcu "D_reviews". Tento stĺpec reprezentuje slovné hodnotenie danej hry. Vyskytuje sa v ňom 8 hodnôt a to **Overwhelmingly, Negative, Very Negative, Negative, Mostly Negative, Mixed, Mostly Positive, Positive, Very Positive, Overwhelmingly positive.** Tieto hodnoty sme premapovali na číselne hodnoty od 0 po 8.

Nasledujúcou úpravou dát bolo vyplnenie prázdnych hodnôt v stĺpcoch. Na vyplnánie prázdnych hodnôt sme použili funkciu **fillna**, ktorá zadaním stĺpcu a hodnoty vyplní všetký prázdne políčka danou hodnotou. Touto úpravoou prebehli nasledujúce stĺpce: **score, languages, publisher_est, developer_est, coming_soon, English, is_single_player, is_multi_player a D_owners**. Stĺpce, ktoré defaultne obsahujú číselné hodnoty sme dopĺnali mediánom všetkých hodnôt v stĺpci. Stĺpce, ktoré obsahovali boolean hodnoty sme dopĺnali boolean hodnotou, ktorá sa prevažne objavovala v stĺpci.

Väčšou úpravou prešiel stĺpec “D_tags”. V tomto stĺpci sa nachádzali Steam tags, ktoré bližšie opisujú žáner, sub-žáner, vizuály, témy, módy a iné vlastnosti hry. Tieto tagy boli zadávané samotnými hráčmi a vlastníkmi hier. Dáta v poličkách boli reprezentované JSON objektom v tvare stringu. Dáta bolo potrebné správne sparsovať. Parsovanie prebiehalo vo viacerých ktorkoch. Ako prvé sme tieto stringové objekty namapovali na python objekty aby sme cez ne vedeli plynule iterovať. Následne sme si vybrali tagy, ktoré chceme vytiahnuť a použiť neskôr. Zvolili sme si týchto pár tagov: **2D, 3D, PvE, PvP, VR**. Lambda funkciou sme prešli objektami a filtrovali sme iba tie atribúty , ktoré sme si zvolili. Po filtrácii sme One Hot Encodli tieto tagy. One Hot Encode nám z každej hodnoty spravil separátny stĺpec a hodnotou 0 alebo 1 sa namapovali políčka podľa toho, či sa v objekte nachádzal nami zvolený tag alebo nie. Stĺpec “D_tags” sme teda týmto spôsobom odstránili a nahradili ho našími piatimi novými stĺpcami.

Niektoré stĺpce obsahovali boolean hodnoty. Tieto hodnoty pre jednoduchšiu manipuláciu sme premapovali na číselné hodnoty. True hodnota bola premapovaná na 1 a False hodnota bola premapovaná na 0. Stĺpce, ktoré prešli takovtou úpravou boli:

'english','is_free','self_published','has_dlc','has_website_linked','has_controller_support','is_single_player','is_multi_player','is_early_access','mature_content','Addictive','Beautiful','Classic','Competitive','Cult Classic','Difficult','Emotional','Emotional','Epic','Funny','Lore-Rich','Masterpiece','Replay Value’,'Short','Well-Written'.

Pár stĺpcov obsahuje hodnoty mimo intervalu 0 až 1. Tým, že vo väčšine stĺpcov sa hodnoty pohybujú presne v tom intervale, stĺpce s inými hodnotami kazia kvalitu výpočtu. Pre tento prípad

je vhodné dáta upravit a premapovať ich do intervalu 0 až 1. Škálovanie dát sme aplikovali pomocou knižnice **mlxtend.preprocessing.** Táto knižnica nám ponúka MinMaxScaling na dáta. MinMaxScaling požaduje dataset a stĺpce , ktoré chceme škálovať. Na túto úpravu sme vybrali stĺpce: **‘D_owners’,'score','D_release_date','languages','D_reviews'.** Teraz sa hodnoty v celom datasete pohybujú v intervale 0 až 1.

	D_owners	score	D_release_date	languages	D_reviews
0	0.000300	0.406250	0.809524	0.000000	0.333333
1	0.000000	1.000000	0.952381	0.000000	0.666667
2	0.004801	0.732530	0.809524	0.000000	0.500000
3	0.000300	0.672566	0.904762	0.285714	0.333333
4	0.000000	1.000000	0.809524	0.000000	0.666667
...	...	...	...	...	...
995	0.000000	0.666667	0.809524	0.000000	0.333333
996	0.000000	0.666667	0.857143	0.250000	0.333333
997	0.000300	0.571429	0.952381	0.000000	0.333333
998	0.000000	0.888889	0.952381	0.178571	0.666667
999	0.000000	1.000000	0.809524	0.000000	0.666667

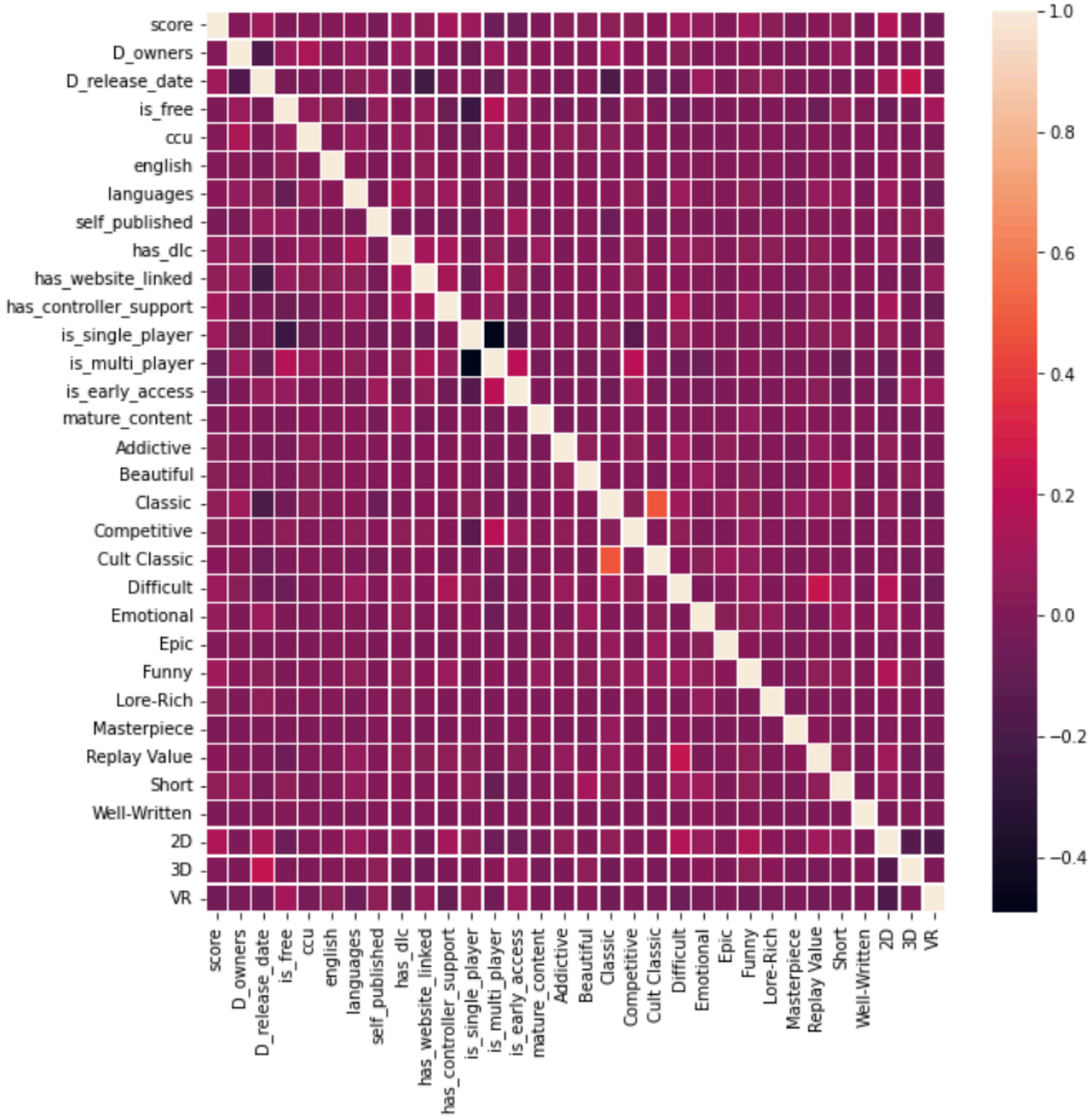
Ako posledné sme dropli stĺpce z tabuľky, ktoré potencionálne nemajú dopad na vyhodnotenie či hra je zadarmo alebo nie. Na dropovanie sme zvolili tieto stĺpce:

’D_appid’,'D_name','D_genre','D_publisher','positive','negative','coming_soon','VYMAZAT_price','D_developer','publisher_est','developer_est'.

Tieto všetký úpravy sme aplikovali aj na testovacie dáta aj na trénovacie dáta.

Korelačná matica

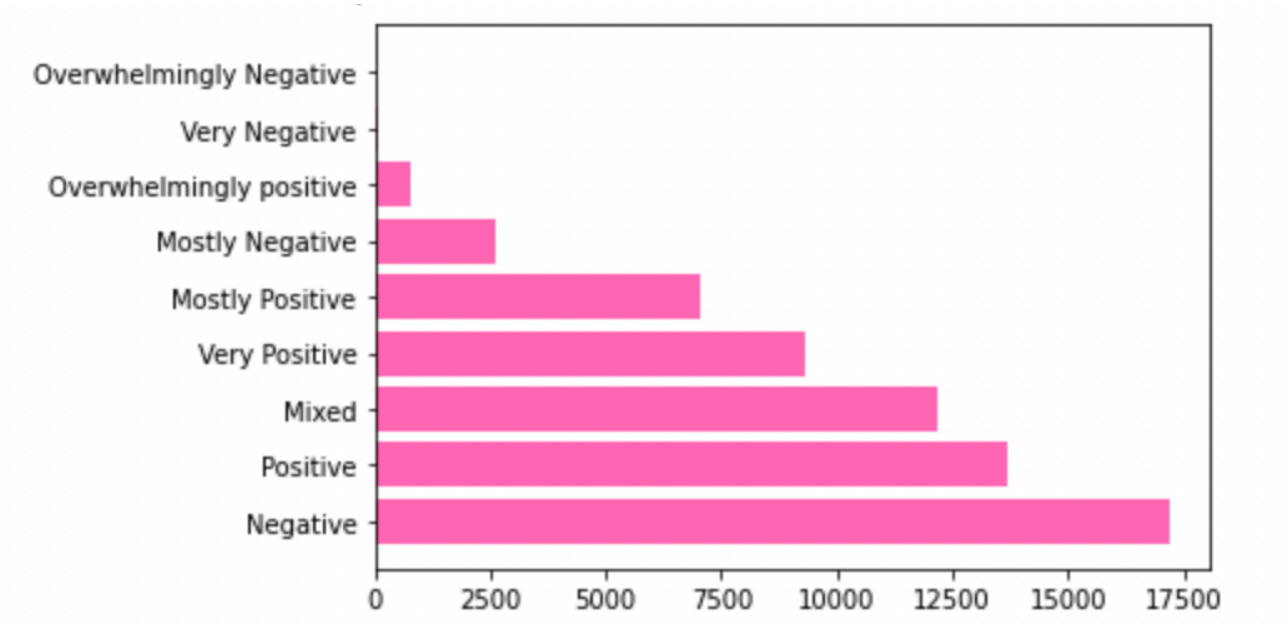
Vďaka koleračnej matici môžeme vizuálne znázorniť závislosti medzi dvoma alebo viacerými premennými. Jednotlivé vzťahy sú podľa prekrývania hodnôt farebne označané. Ak majú dva stĺpce spolu spoločných veľa hodnotť v rovnakých riadkoch tak je štvorec svetlejší. V opačnom prípade štvorec tmavne. Ako môžeme vidieť hlavná diagonála je celá svetlá, čo značí že hodnoty sa prekrývajú. Ako môžeme napríklad vidieť stĺpec is_multi_plaeyer a stĺpec is_single_player majú prienik čierny štvorec. Tieto hodnoty su samozrejme vždy odlišné, pretože dáta s hrami majú vždy protichodnú hodnotu v týchto stĺpcoch.



Analýza datasetu cez EDA

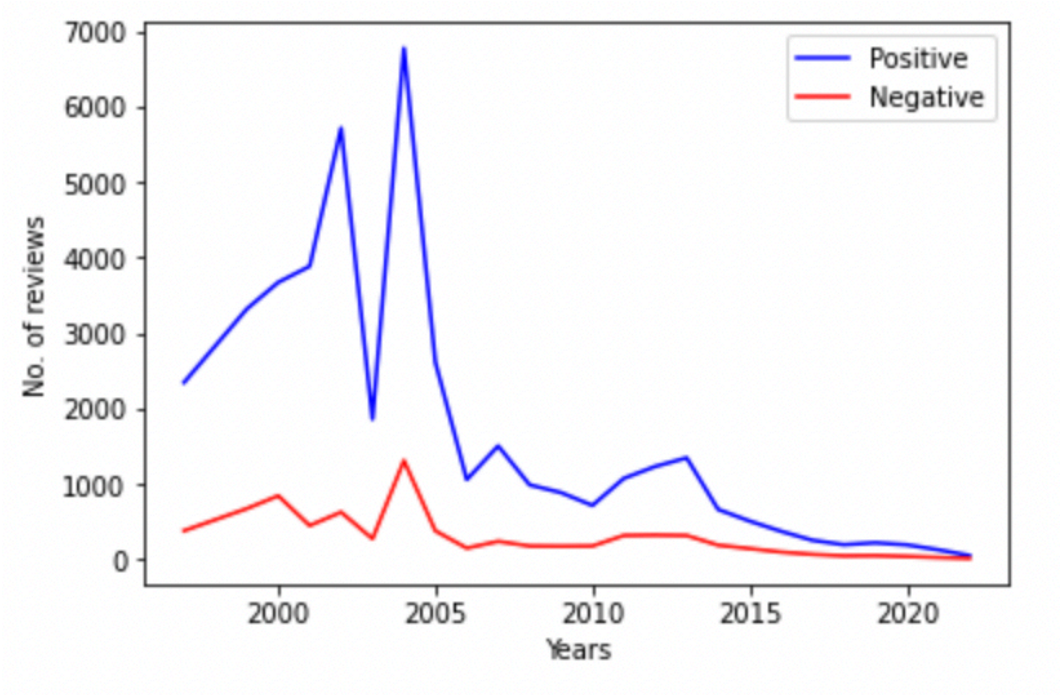
Pre analýzu datasetu sme si pripravili 6 grafov, ktoré zobrazujú súvislosti medzi danými stĺpcami alebo vysvetlenie účelu stĺpca. Na vizualizáciu sme použili histogramy, chart grafy, stĺpcové grafy a koláčový graf.

1. Hodnotenia hier



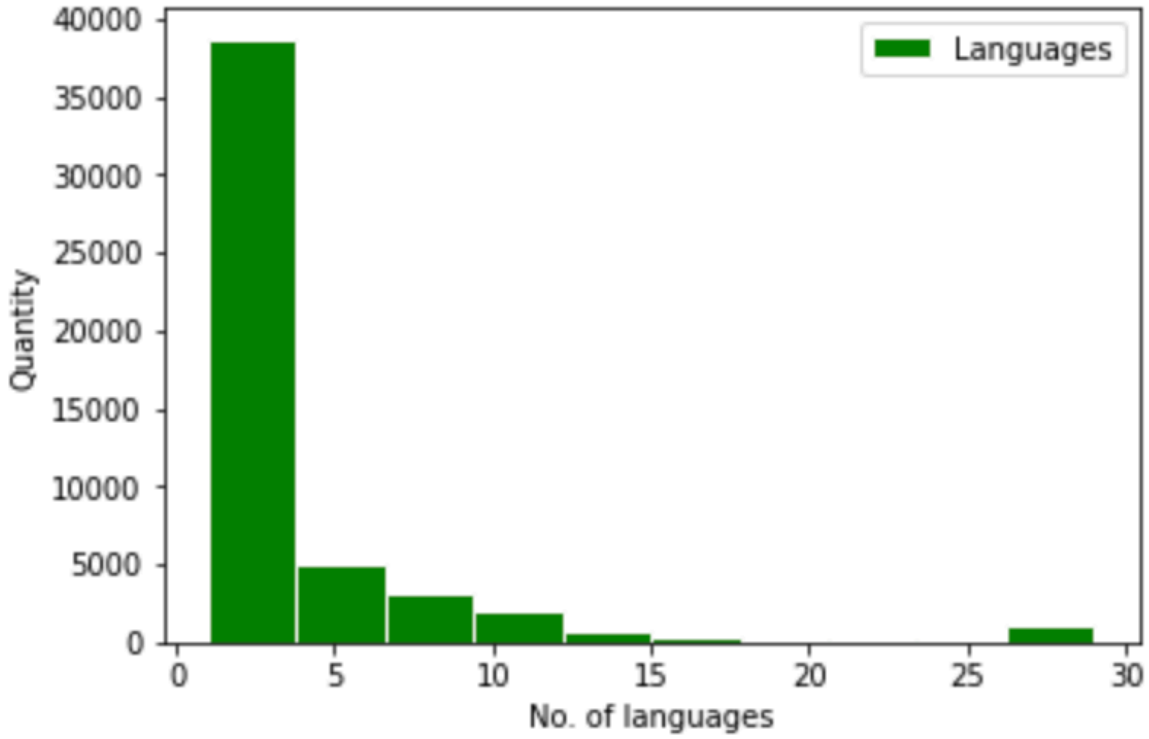
Na stĺpcovom grafe môžeme vidiet slovné hodnotenia hier na platforme Steam. Na x-ovej osi sú početnosti daných hodnotení. Na y-ovej osi môžeme vidieť slovné hodnotenia. Z grafu vyplýva, že žiadna hra nemá celkové hodnotie Overwhelmingly Negative a Very Negative ale na druhej strane prevládajú Negative hodnotenia. Po prehlaidnutí grafu môžeme usúdiť, že hodnotenia sú vyrovnané pretože, máme vela dobrých hodnotení ale stĺpec Negative prevláda a tento priemer dobrých hodnotení znižuje.

2. Pozitívne a negatívne hodnotenia



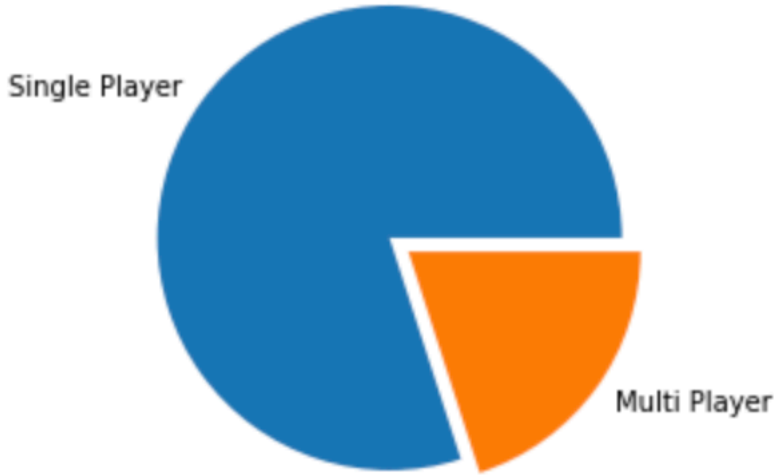
Na chart grafe môžeme vidieť počet pozitívnych a negativných hodnotení v časovom priebehu. Modrá čiara reprezentuje pozitívne hodnotenia a červená čiara reprezentuje negatívne hodotenia. Na x-ovej osi vidíme roky a na y-osi vidíme počet hodnotení. Z grafu vyplýva, že prirodzene hry, ktoré sú staršie majú viac hodnotení. Taktiež vidíme, že o mnoho viac je pozitívnych hodnotení hier ako negatívnych. Pri roku 2020 sa nám čiary stretávajú z čoho vyplýva, že novšie hry majú výrazne menej hodnotení ako hry napríklad v rokoch 2002 až 2005.

3. Jazyky hier



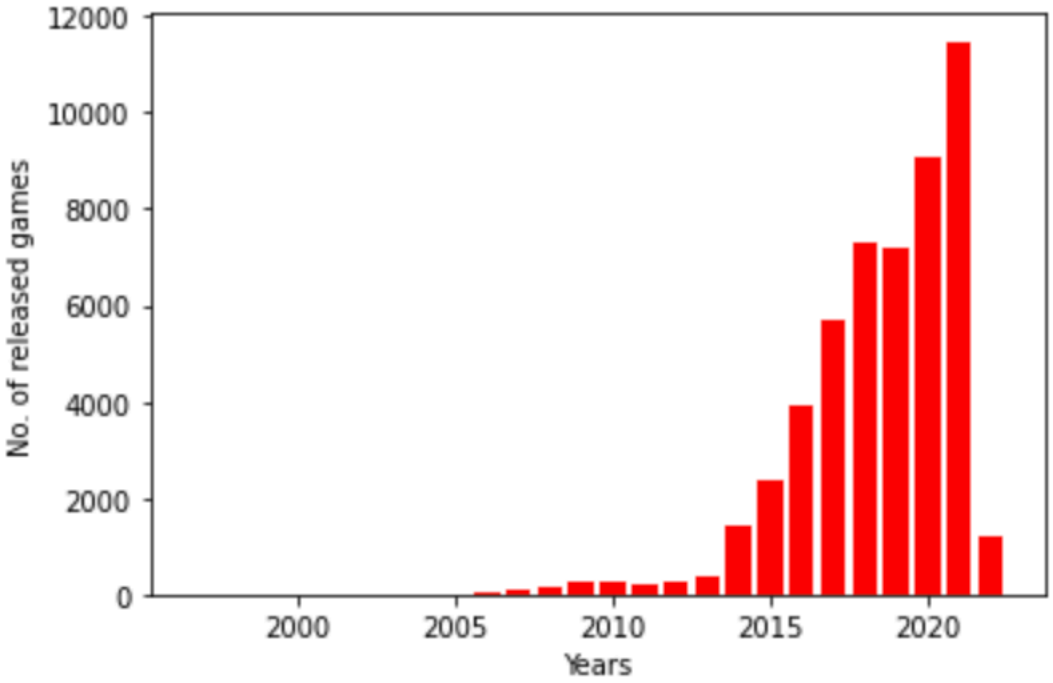
Na histograme môžeme vidieť celkový počet jazykov hier. Na x-ovej osi vidíme počet jazykov a na y-ovej osi vidíme početnosť hier s počtom jazykov. Z histogramu vyplýva, že drtivá väčšina hier je preložená do 1 až 4 jazykov. Taktiež môžeme vidieť, že zlomok hier je preložený do 27 až 29 jazykov.

4. Single player a Multi player hry



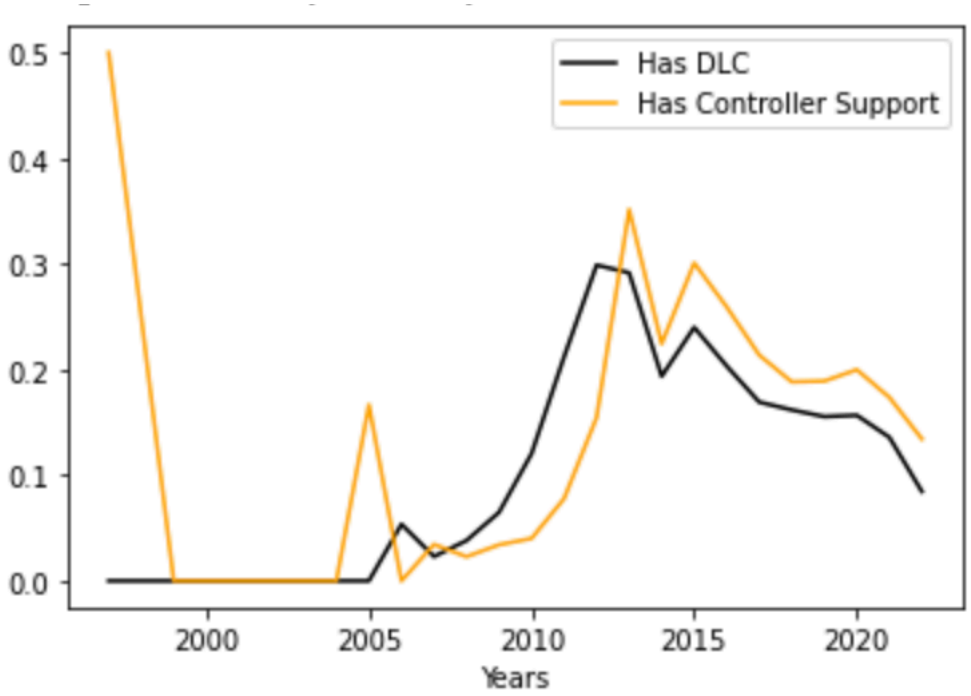
Na koláčovom grafe môžeme vidieť pomer Single player hre ku Multi player hrám. Modrá časť grafu reprezentuje množinu single player hier a oranžová časť grafu reprezentuje množinu multi player hier. Z grafu vyplýva, že vyše trištvrtina hier sú single player hry, čiže kampaňové hry. Tento pomer je prirodzený, pretože multi player hry začali byť populárne a produkované až o mnoho nesôr než single player hry.

5. Dátum vydania hier



Na stĺpcovom grafe môžeme vidieť množstvo vydaných hier v daných rokoch. Na x-ovej osi vidíme roky a na y-ovej osi vidíme početnosť vydaných hier. Rastúcou popularitou počítačovych hier a vylepšovaním enginov na vytváranie hier počas prvého desatročia prirodzene rástol počet hier. Z grafu vyplýva, že po súčastnosť počet hier exponensionálne rástol. Až na posledný stĺpec v grafe, ktorý reprezentuje súčasný rok.

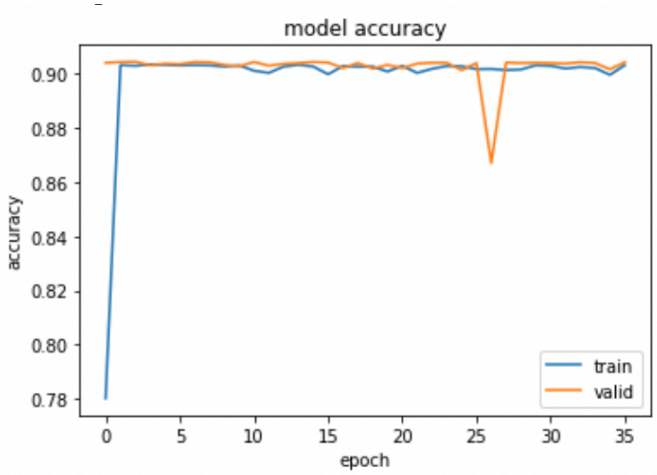
6. DLC a podpora ovládačov

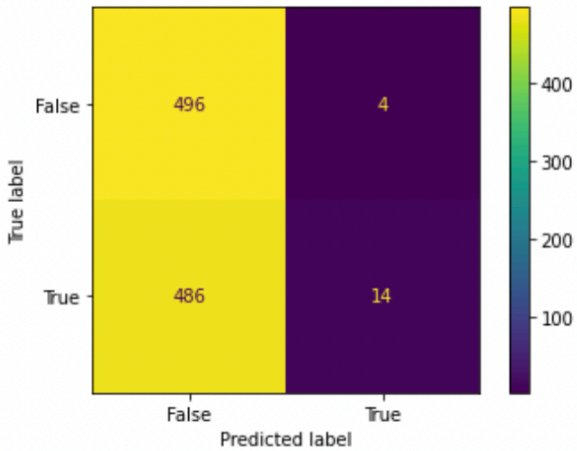


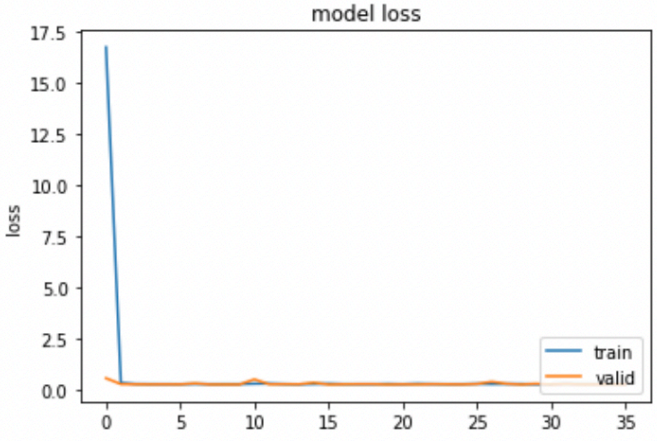
Na grafe môžeme vidieť dve čiary reprezentujúce DLC a podporu ovládačov počas nasledovných rokov. Čierna čiara reprezentuje hry, ktoré majú dodatočné prídavky k hre, DLC. Žltá čiara reprezentuje hry, ktoré maju podporu pre externé ovládače. Na x-ovej osi vidíme roky a na y-ovej osi vidíme hodnoty od 0 po 1, ktoré reprezentujú boolean hodnoty, či hra ponúka taketo rozšírenie. Z grafu vyplýva, že najviac dodatočných prídavkov k hre bolo vydávaných v roku 2010. Čím staršia a populárnejšia hra je, tým viac DLC je pridávaných do hry. Z toho vyplýva, že nové hry majú prirodzene menej prídavkov. Podpora pre ovládače bola zavedená počas eŕi herných konzoli, ktorá začala cca v roku 2005. Od toho roku môžeme vidieť vysoký nárast hier s podporou pre externé ovládače.

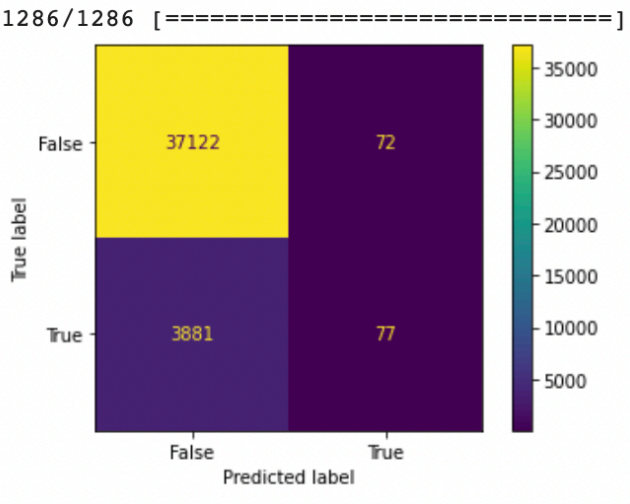
Trénovanie neurónovej siete

Prvým bodom pri trénovaní našej neurónovej siete bolo rozdelenie trénovacícch dát do dvoch skupín. 80% dát sme ponechali na trénovanie a zvyšných 20% dát sme vyčlenili ako validačné dáta. Následne sme si vyčlenili input a output dáta. Input dáta sú dáta, na ktorých sa bude trénovať bez stĺpca “is_free”. Output dáta sú dáta, ktoré pozostávajú len z jedného stĺpca “is_free”. Týmto spôsobom sme vyčlenili trénovacie, testovacie i validačné dáta. Po vytvorení modelu sme pridávali vrstvy do modelu. Pridali sme dve aktivačné vrstvy typu **relo** s 12 a 8 neurónmi a jednu typu **sigmoid** s 1 neurónom. Použili sme optimalizačný algoritmus **adam**, ktorý upravuje váhy hodnôt za behu. Ako kriteriálna funkcia pri trénovaní sme použili **binary_crossentropy**. Taktiež sme nastavili **early_stopping** na siete pre elimináciu rizika pretrénovania siete. Trénovanie sme nastavili na 250 epoch pri batch_size 200.

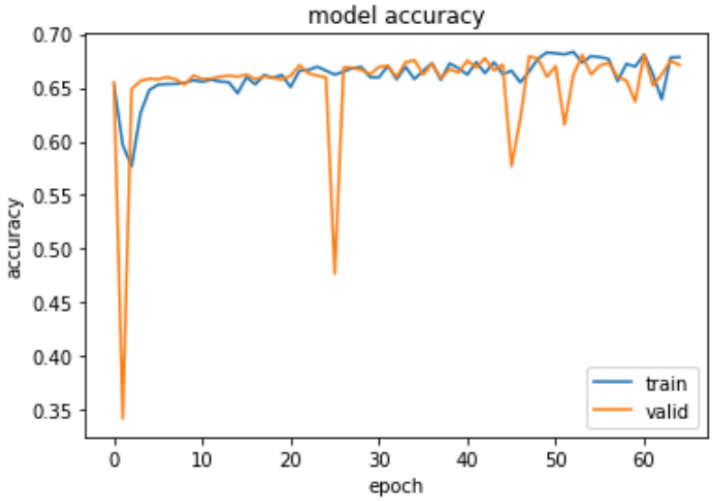


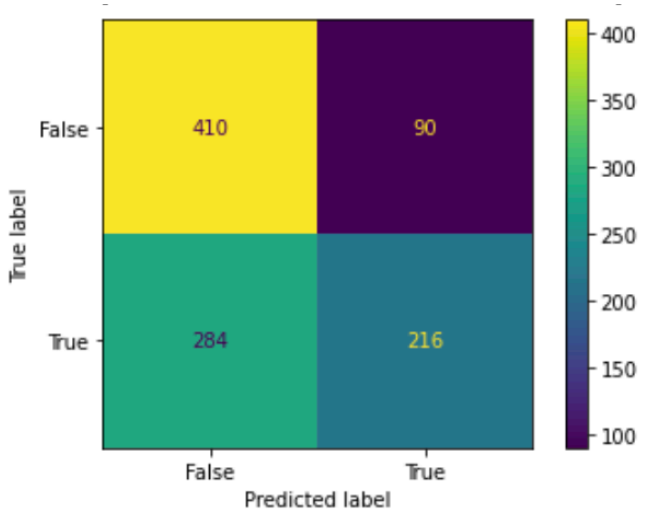


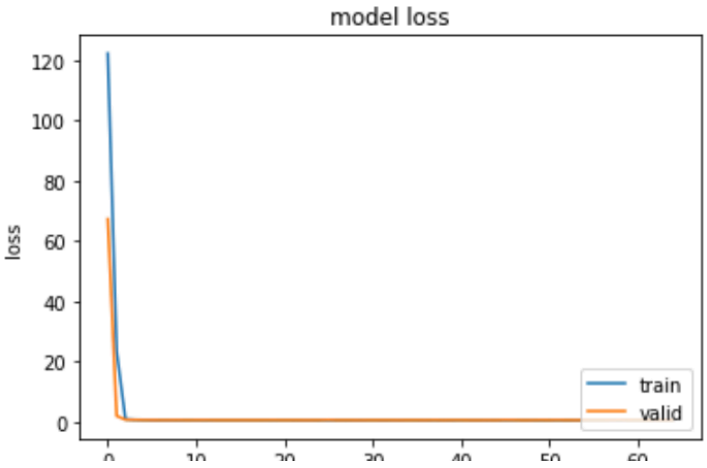


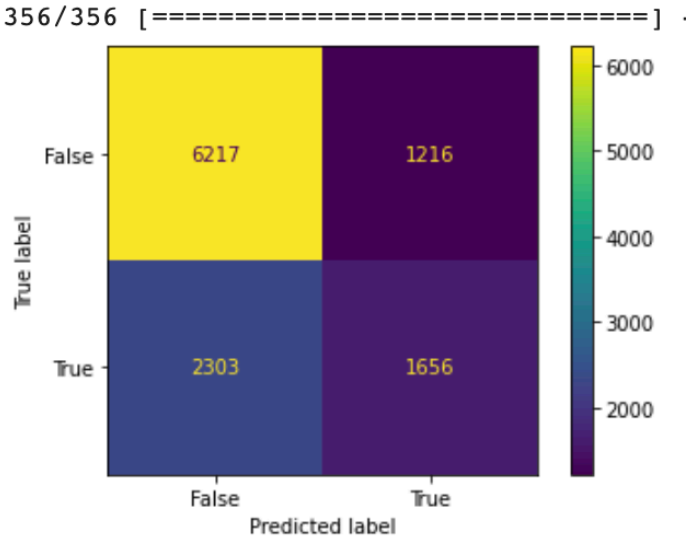


Celková presnosť vytrénovanej siete vyšla na 51%. Ako môžeme vidieť na konfúznych maticiach takmer všetký hry, ktoré mali byť zadarmo, naša sieť zhodnotila ako platené hry. Naopak takmer všektý hry, ktoré sú platené, zhodnotila naša sieť správne. Po preskúmaní dát sme prišli na nezrovnalosti v dátovej množine. Platených hier je 5 násobne viac ako hier zadarmo. Preto najlepším riešením bolo dropnúť istú časť množiny dát, kde sú hry platené.



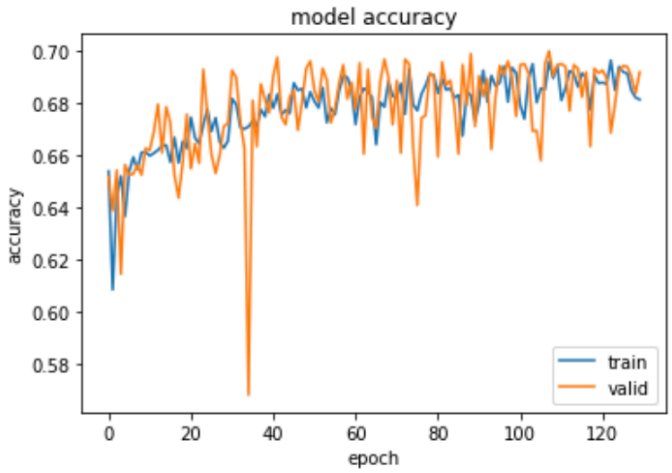


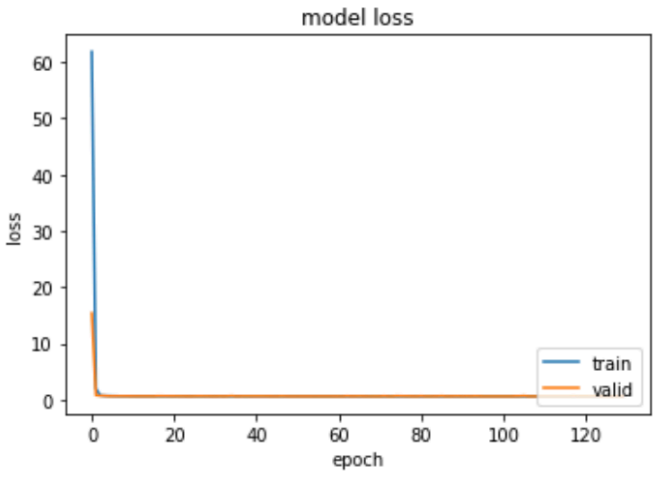


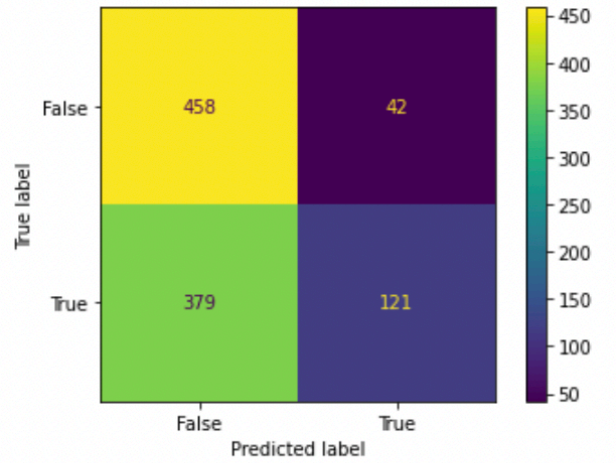


Po úprave dát boli výsledky o mnoho lepšie. Úspešnosť stúpla na 63% a na konfúznej matici môžeme vidieť, že naša sieť lepšie hodnotí cenu hier.

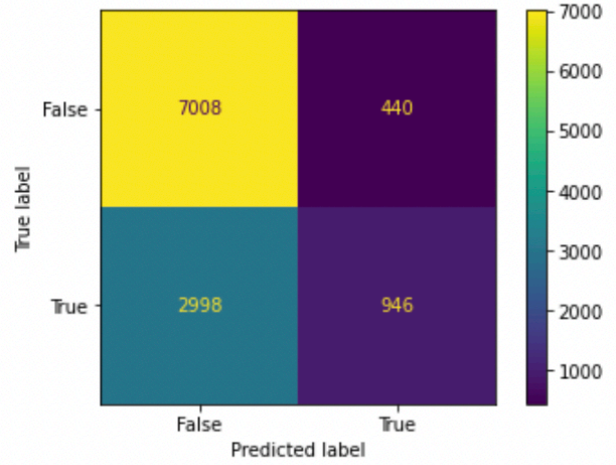
Ďalším krokom bolo experimentovanie s množstvom epoch a batch_sizeom. Upravili sme množstvo epoch z 250 na 150 a batch_size z 200 na 100.



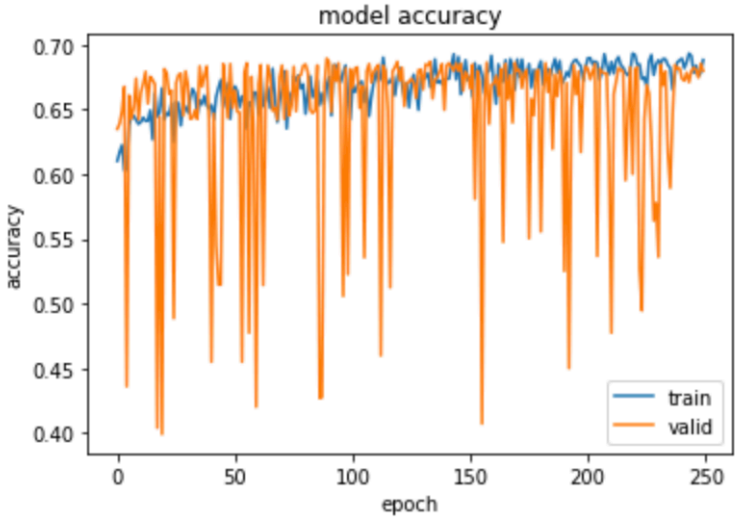


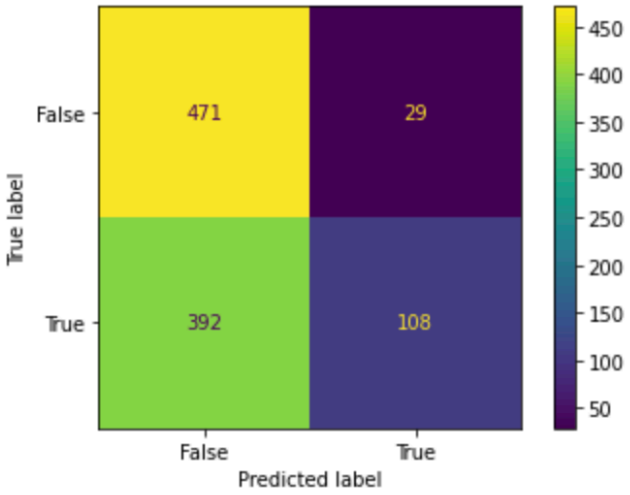


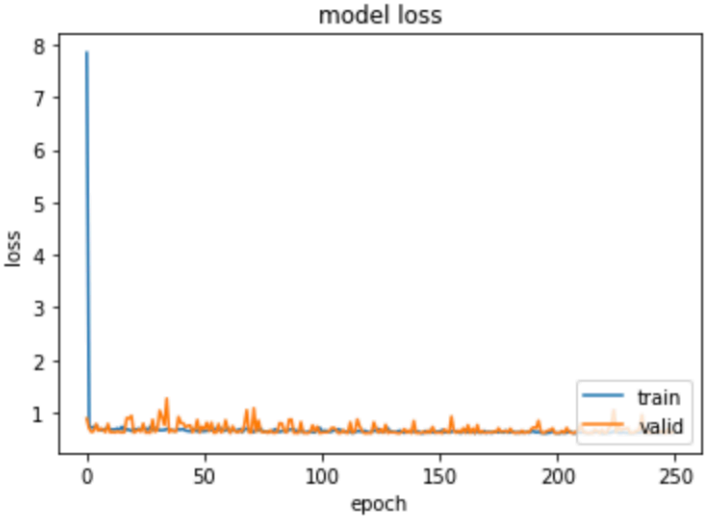
356/356 [==============================]



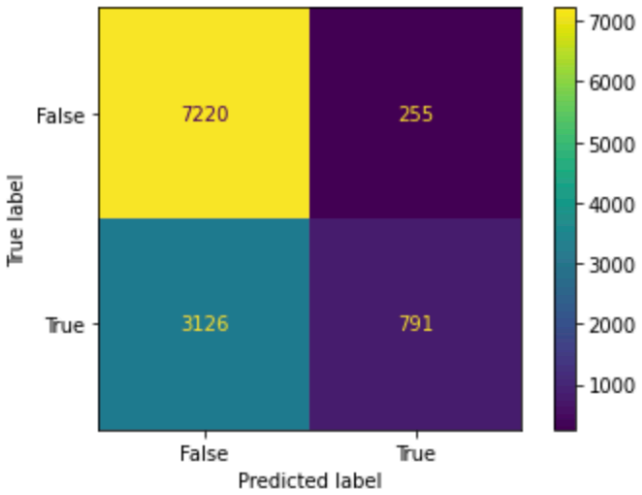
Bohužiaľ presnosť našej siete klesla na 57.9%. Pri znížení počtu sme nemali dobré výsledky. Ďalším experimentom bolo zvýšenie počtu epóch na 250 a batch_size na 50. A taktiež sme skúsili odstrániť early_stopping funkciu.



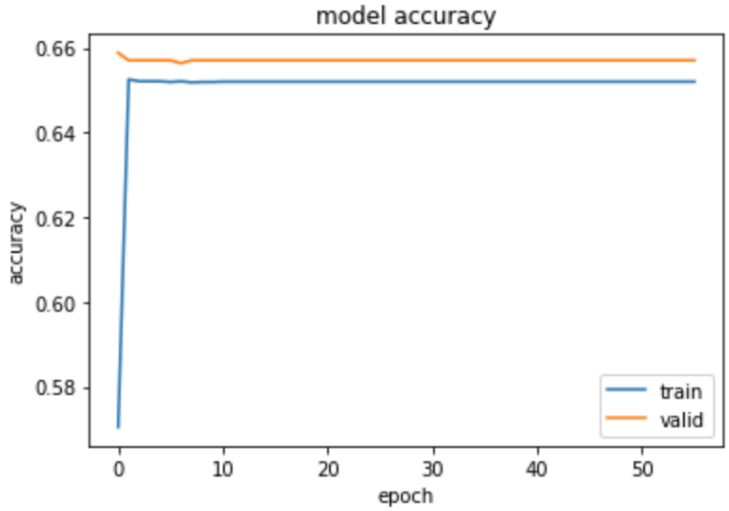


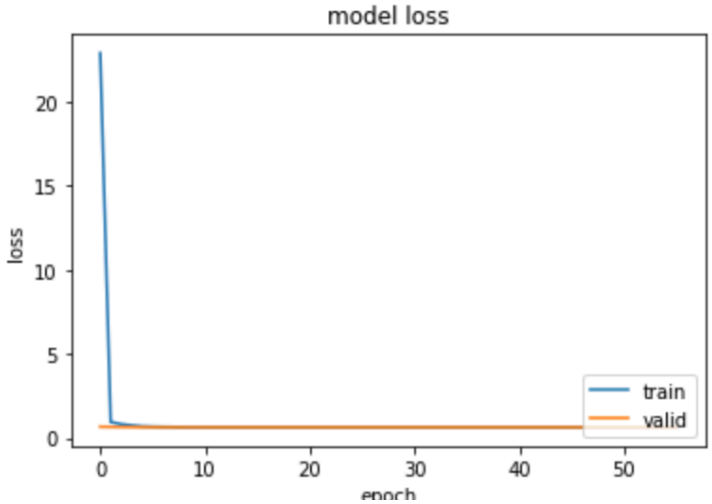


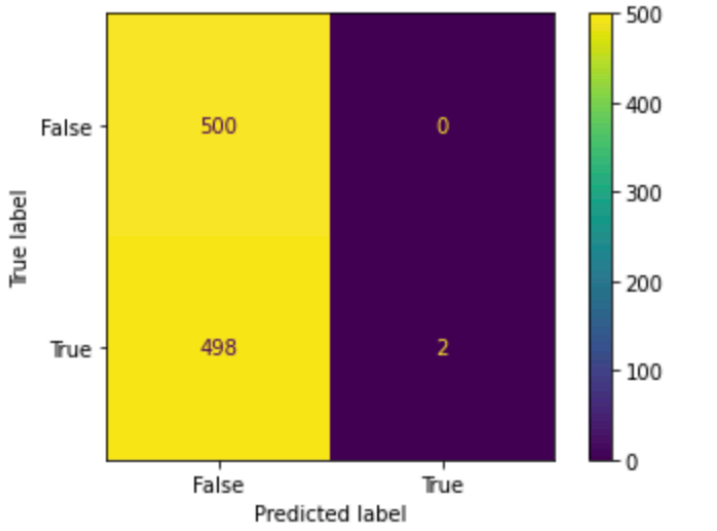
356/356 [==============================] -



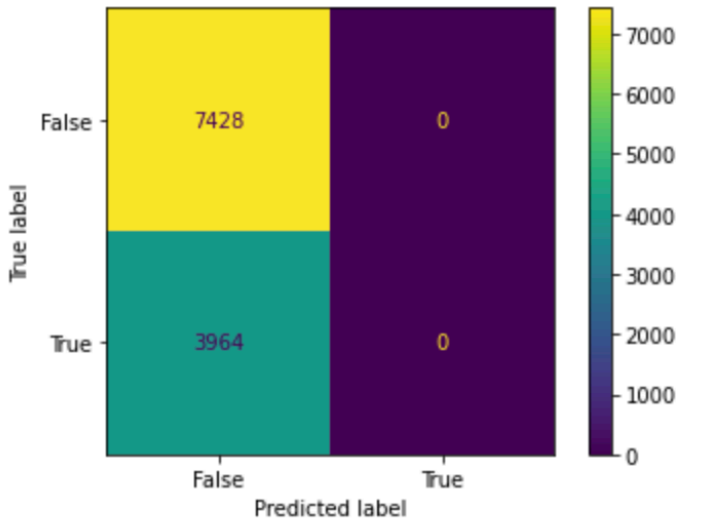
Prekvapivo výsledky presnosti siete sú totožné s predošlým experimentom a to 57.9%. Ďalej sme skúsili ponechať počet epóch aj batch_size ale nastavili sme znova early_stoping funkcie. Je možné, že tieto parametre boli lepšie ako predošlé ale kvôli tomu, že sme odstránili early_stoping funkciu sa nam sieť pretrenovala.







356/356 [==============================] -



Early_stoping zapríčínil, že sieť sa prestala trémovať pri 56. epoche a vypočítalo presnosť 50.2%. Týmto smerom nebolo vhodné pokračovat, preto sme sa rozhodli vrátiť k hodnotám, ktoré nám trénovali sieť najlepšie. Počet epóch 250 a batch_size 200. Výsledky siete sme už uviedli, presnosť presiahla 60%.

Pri experimentovaní s počtom neurónov vo vrstvách sme neprišli k lepším výsledkom, práveže naopak. Preto sme sa rozhodli ostať s pôvodnými počtami.

Záver

Experimenti nám moc nevychádzali a prvotné parametre boli podľa štatistík najlepšie. Najvyššiu presnosť vytrénovanej neurónovej siete sme dosiahli 63%. Túto presnosť sme dosiahli hlavne kvôli dropnutiu dát, ktoré kazili kvalitu výpočtu. V datasete sa nachádzali väčšinou dáta s platenými hrami a neurónova sieť preto vyhodnocovala väčšinu hier ako platených a klesala presnosť.