Sales Maximization Model: Steam

Authors: Kelly Chen, Joshua Petrikat, Justin Tran, Paul Yokota, Bryan Yu
Group 7 (Interactix)
STAT167

Objectives and Questions of Interest

Primary Objectives:

Identify the most important factors influencing high sales (over 20,000) in Steam games.

Secondary Objective

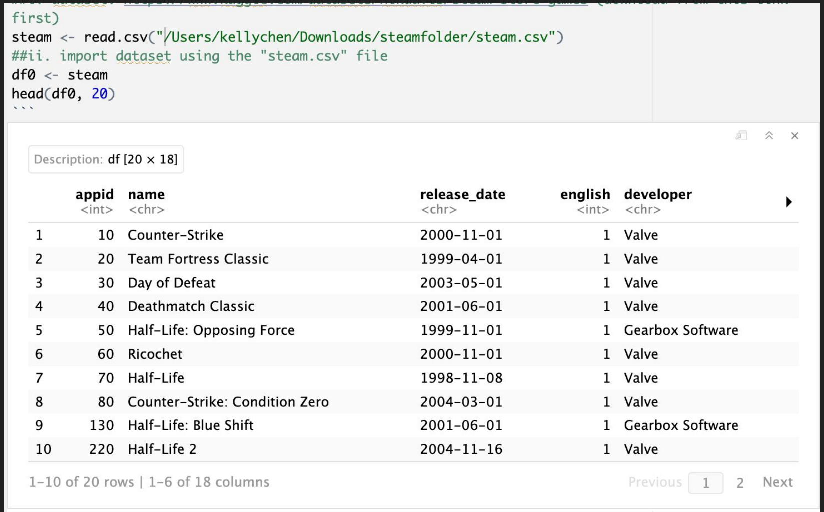
1) Clean and prepare the Steam game data (df) for effective machine learning model training.
2) Develop a machine learning model to predict with high accuracy whether a Steam game will achieve over 20,000 sales.
3) Compare the performance of different machine learning models (logistic regression, random forest, and support vector machines) for predicting high sales.

Questions of Interest:

1) Does being a renowned publisher increase the chances of a game having high sales?
2) What genres (tags) corresponded to the best reception?
3) Does making your game available on all three operating systems (mac, windows, and linux) increase the probability of high sales (over 20k)?

The Data

- Contains 27,075 observations (titles of games on Steam) of 18 variables
- Each observation is a game listed and sold on Steam between 1997 and 2019.



```
first)
steam <- read.csv("/Users/kellychen/Downloads/steamfolder/steam.csv")
##ii. import dataset using the "steam.csv" file
df0 <- steam
head(df0, 20)
```

Description: df [20 × 18]

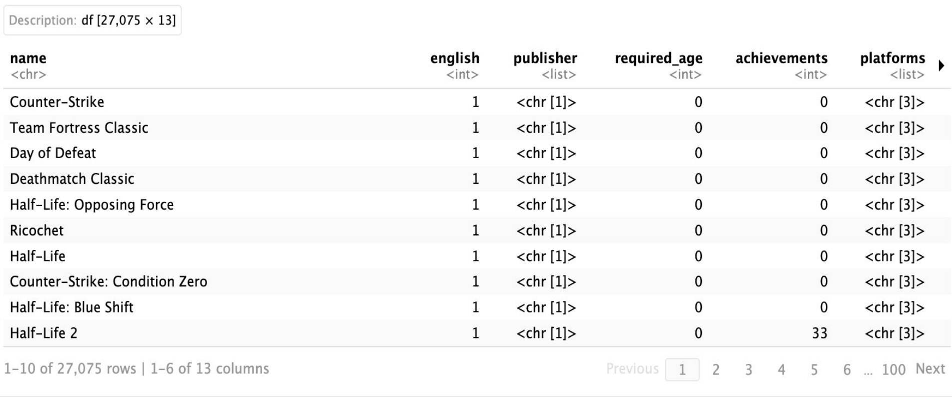
	appid <int>	name <chr>	release_date <chr>	english <int>	developer <chr>
1	10	Counter-Strike	2000-11-01	1	Valve
2	20	Team Fortress Classic	1999-04-01	1	Valve
3	30	Day of Defeat	2003-05-01	1	Valve
4	40	Deathmatch Classic	2001-06-01	1	Valve
5	50	Half-Life: Opposing Force	1999-11-01	1	Gearbox Software
6	60	Ricochet	2000-11-01	1	Valve
7	70	Half-Life	1998-11-08	1	Valve
8	80	Counter-Strike: Condition Zero	2004-03-01	1	Valve
9	130	Half-Life: Blue Shift	2001-06-01	1	Gearbox Software
10	220	Half-Life 2	2004-11-16	1	Valve

1-10 of 20 rows | 1-6 of 18 columns

Source: https://www.kaggle.com/datasets/nikdavis/steam-store-games

Cleaning

- Created ‘reception’
- Kept year from release date
- Changed tags to factored variables
- Transformed owners into bins
- Determined whether a game was sold by a top-grossing publisher or not

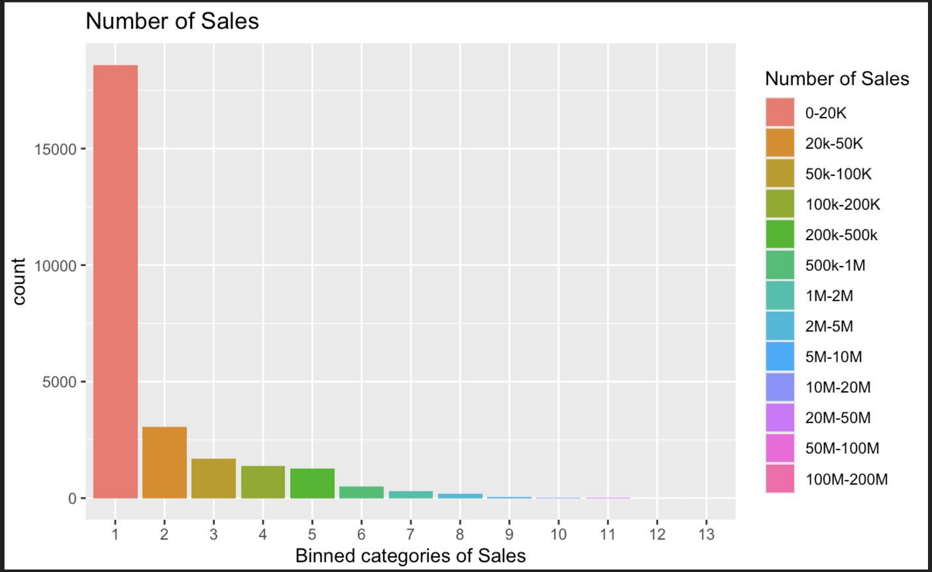


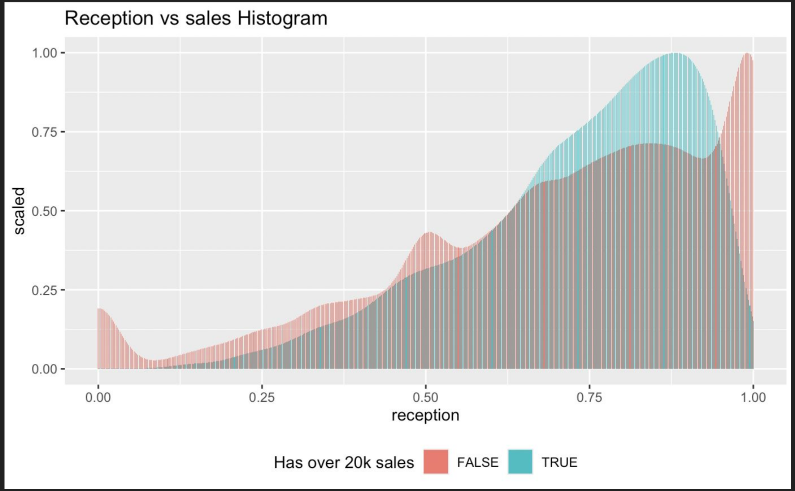
Description: df [27,075 × 13]

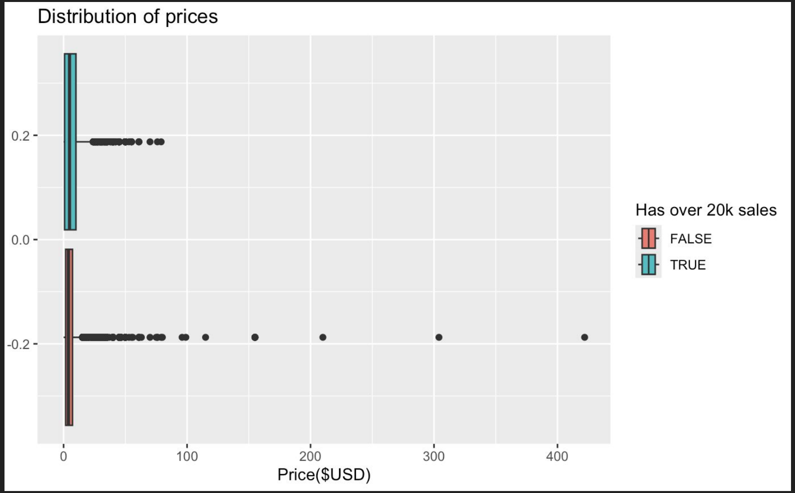
name <chr>	english <int>	publisher <list>	required_age <int>	achievements <int>	platforms <list>
Counter-Strike	1	<chr [1]>	0	0	<chr [3]>
Team Fortress Classic	1	<chr [1]>	0	0	<chr [3]>
Day of Defeat	1	<chr [1]>	0	0	<chr [3]>
Deathmatch Classic	1	<chr [1]>	0	0	<chr [3]>
Half-Life: Opposing Force	1	<chr [1]>	0	0	<chr [3]>
Ricochet	1	<chr [1]>	0	0	<chr [3]>
Half-Life	1	<chr [1]>	0	0	<chr [3]>
Counter-Strike: Condition Zero	1	<chr [1]>	0	0	<chr [3]>
Half-Life: Blue Shift	1	<chr [1]>	0	0	<chr [3]>
Half-Life 2	1	<chr [1]>	0	33	<chr [3]>

1-10 of 27,075 rows | 1-6 of 13 columns

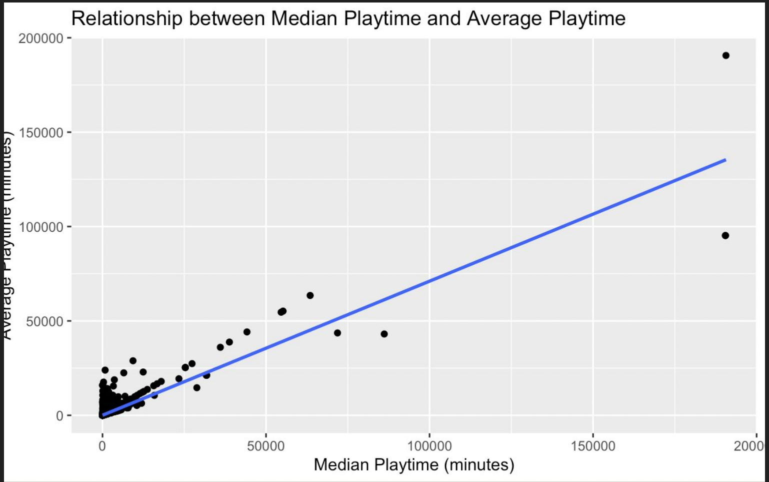
Exploratory Data Analysis

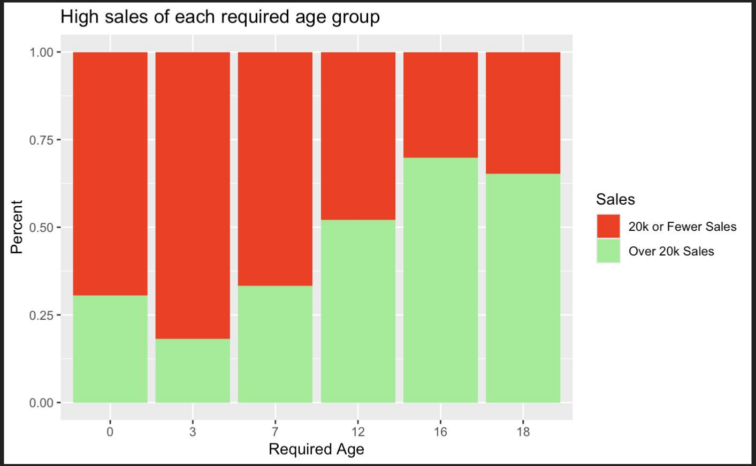


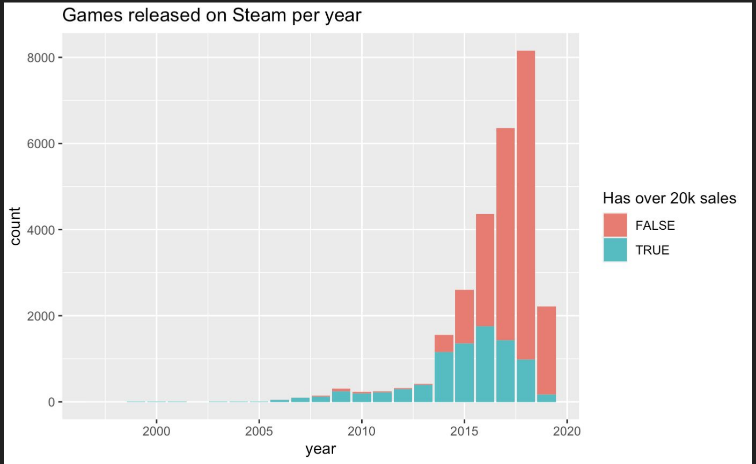




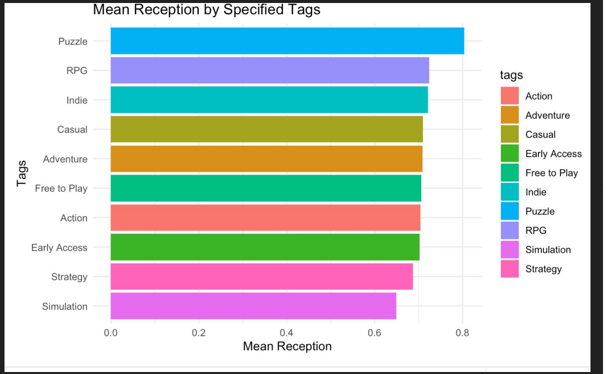
Exploratory Data Analysis

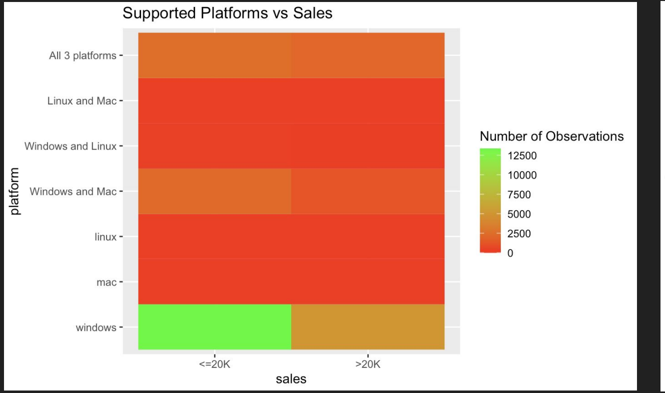


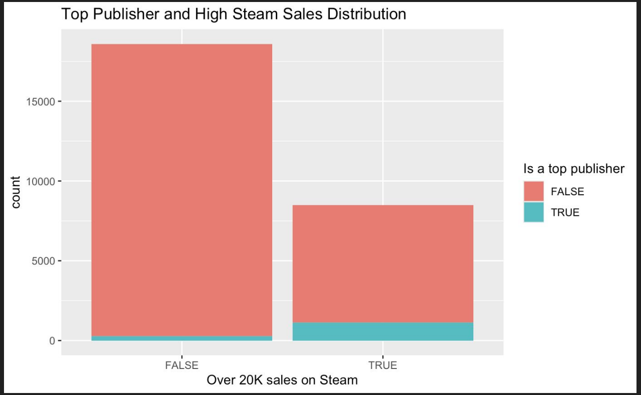




Exploratory Data Analysis - Questions of Interest







Which genres (tags) corresponded to the best reception.

- Puzzle
- RPG
- Indie

The heatmap shows a very large amount of games that are only supported on windows.

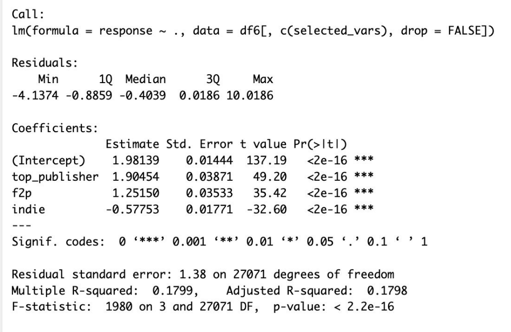
Do top publishers produce more high-selling games?

- Yes

Forward Selection Method

What are the 3 most important variables for increasing sales?

- Be a top publisher
- Free to play
- Not in the Indie genre

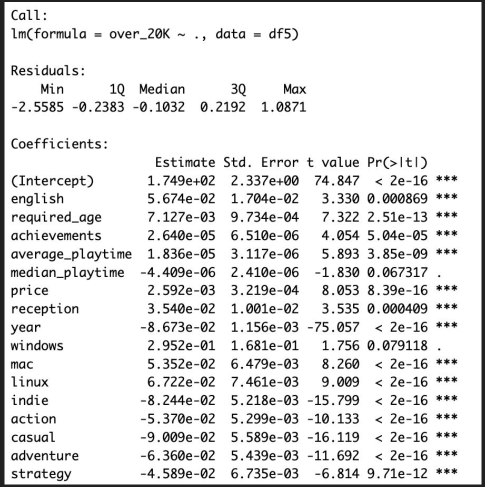


```
Call:
lm(formula = response ~ ., data = df6[, c(selected_vars), drop = FALSE])

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1374 -0.8859 -0.4039  0.0186 10.0186 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.98139    0.01444  137.19   <2e-16 ***
top_publisher  1.90454    0.03871   49.20   <2e-16 ***
f2p            1.25150    0.03533   35.42   <2e-16 ***
indie         -0.57753    0.01771  -32.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.38 on 27071 degrees of freedom
Multiple R-squared:  0.1799,	Adjusted R-squared:  0.1798 
F-statistic:  1980 on 3 and 27071 DF,  p-value: < 2.2e-16
```

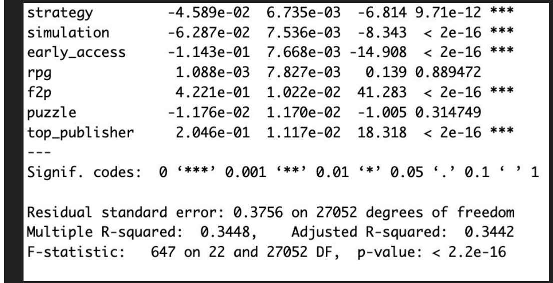
Full Linear Regression

Linear model is not a good predictor of sales.

- Low R^2 value: 0.3442
- High residual standard error in the context of our model: 0.3756



```
Call:
lm(formula = over_20K ~ ., data = df5)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5585 -0.2383 -0.1032  0.2192  1.0871 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.749e+02  2.337e+00  74.847  < 2e-16 ***
english           5.674e-02  1.704e-02   3.330 0.000869 ***
required_age      7.127e-03  9.734e-04   7.322 2.51e-13 ***
achievements      2.640e-05  6.510e-06   4.054 5.04e-05 ***
average_playtime  1.836e-05  3.117e-06   5.893 3.85e-09 ***
median_playtime  -4.409e-06  2.410e-06  -1.830 0.067317 .  
price             2.592e-03  3.219e-04   8.053 8.39e-16 ***
reception         3.540e-02  1.001e-02   3.535 0.000409 ***
year             -8.673e-02  1.156e-03 -75.057  < 2e-16 ***
windows           2.952e-01  1.681e-01   1.756 0.079118 .  
mac               5.352e-02  6.479e-03   8.260  < 2e-16 ***
linux             6.722e-02  7.461e-03   9.009  < 2e-16 ***
indie            -8.244e-02  5.218e-03 -15.799  < 2e-16 ***
action           -5.370e-02  5.299e-03 -10.133  < 2e-16 ***
casual           -9.009e-02  5.589e-03 -16.119  < 2e-16 ***
adventure        -6.360e-02  5.439e-03 -11.692  < 2e-16 ***
strategy         -4.589e-02  6.735e-03  -6.814 9.71e-12 ***
```



```
strategy         -4.589e-02  6.735e-03  -6.814 9.71e-12 ***
simulation       -6.287e-02  7.536e-03  -8.343  < 2e-16 ***
early_access     -1.143e-01  7.668e-03 -14.908  < 2e-16 ***
rpg               1.088e-03  7.827e-03   0.139 0.889472    
f2p               4.221e-01  1.022e-02  41.283  < 2e-16 ***
puzzle           -1.176e-02  1.170e-02  -1.005 0.314749    
top_publisher     2.046e-01  1.117e-02  18.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3756 on 27052 degrees of freedom
Multiple R-squared:  0.3448,	Adjusted R-squared:  0.3442 
F-statistic:   647 on 22 and 27052 DF,  p-value: < 2.2e-16
```

Random Forest Model

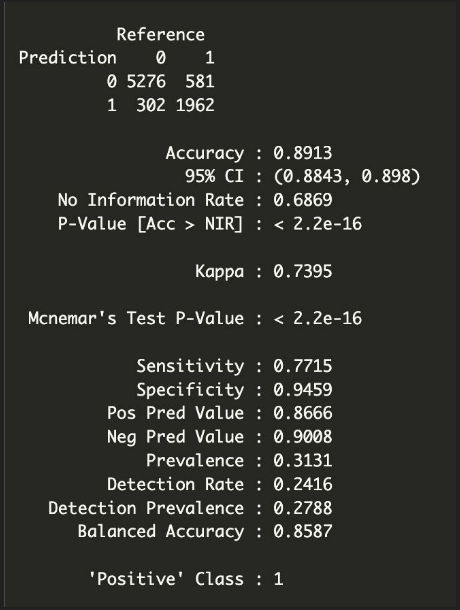
Notable statistics

Accuracy = 89.13%

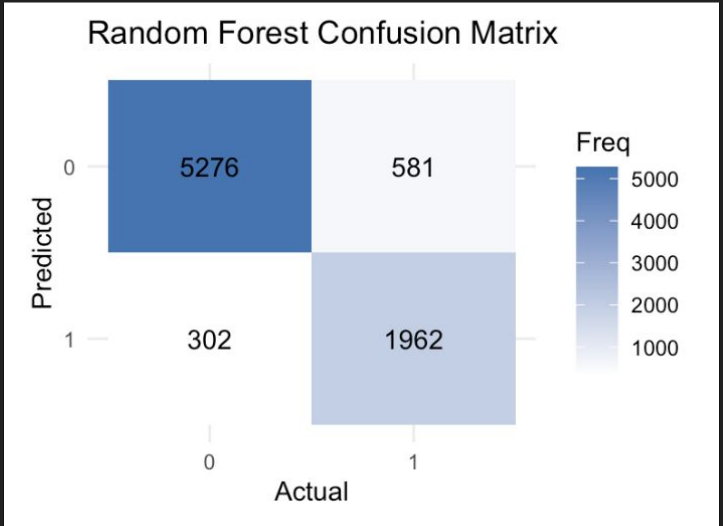
Sensitivity (Recall): 77.15%

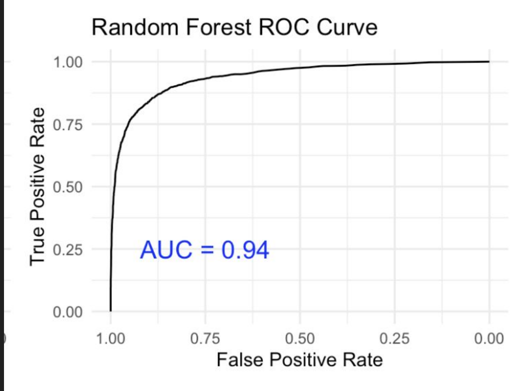
Specificity: 94.59%

AUC (Area Under the ROC Curve): 0.94



```
          Reference
Prediction    0    1
         0 5276  581
         1  302 1962

               Accuracy : 0.8913
                 95% CI : (0.8843, 0.898)
    No Information Rate : 0.6869
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7395

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7715
            Specificity : 0.9459
         Pos Pred Value : 0.8666
         Neg Pred Value : 0.9008
             Prevalence : 0.3131
         Detection Rate : 0.2416
   Detection Prevalence : 0.2788
      Balanced Accuracy : 0.8587

       'Positive' Class : 1
```





Supported Vector Machine (SVM) Model

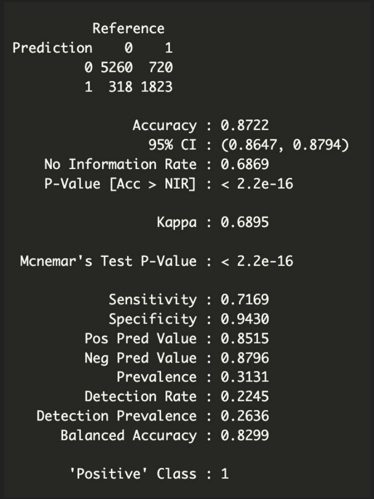
Notable statistics

Accuracy = 87.22%

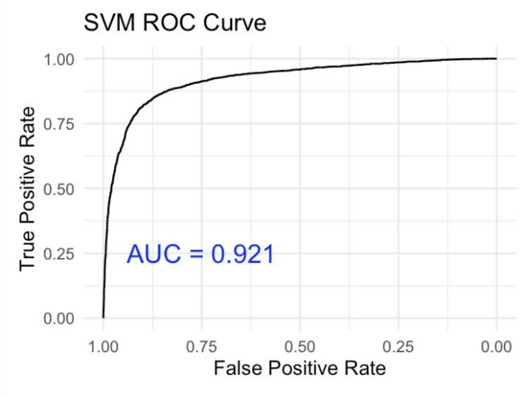
Sensitivity (Recall): 71.69%

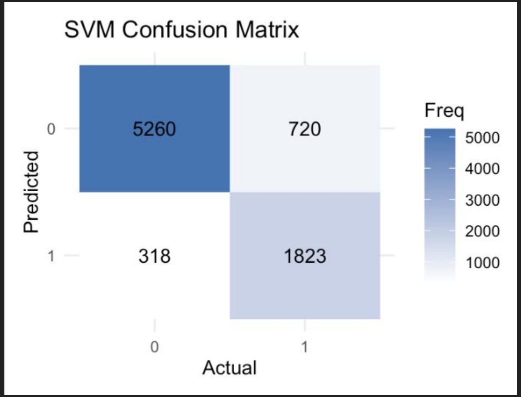
Specificity: 94.3%

AUC (Area Under the ROC Curve): 0.921



```
          Reference
Prediction    0    1
         0 5260  720
         1  318 1823

               Accuracy : 0.8722
                 95% CI : (0.8647, 0.8794)
    No Information Rate : 0.6869
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6895

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7169
            Specificity : 0.9430
         Pos Pred Value : 0.8515
         Neg Pred Value : 0.8796
             Prevalence : 0.3131
         Detection Rate : 0.2245
   Detection Prevalence : 0.2636
      Balanced Accuracy : 0.8299

       'Positive' Class : 1
```





Logistic Model

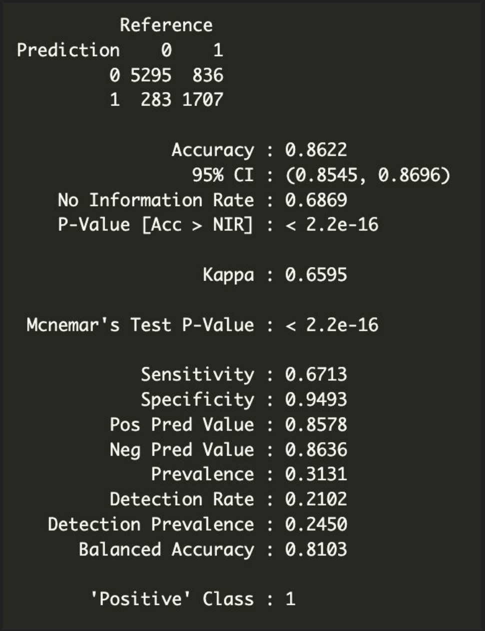
Notable statistics

Accuracy = 86.22%

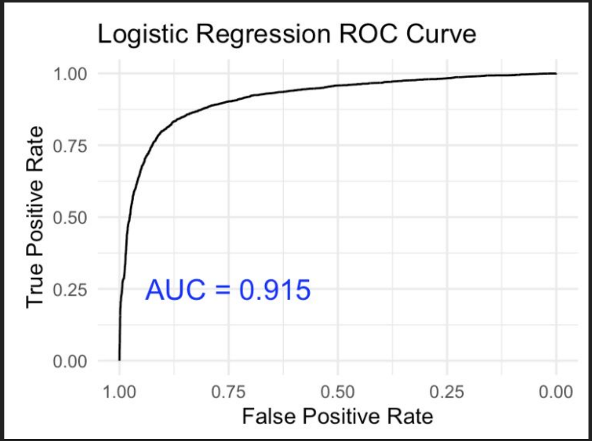
Sensitivity (Recall): 67.13%

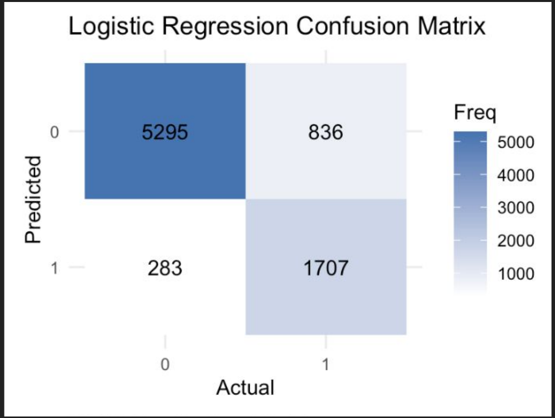
Specificity: 94.93%

AUC (Area Under the ROC Curve): 0.915



```
          Reference
Prediction    0    1
         0 5295  836
         1  283 1707

               Accuracy : 0.8622
                 95% CI : (0.8545, 0.8696)
    No Information Rate : 0.6869
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6595

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6713
            Specificity : 0.9493
         Pos Pred Value : 0.8578
         Neg Pred Value : 0.8636
             Prevalence : 0.3131
         Detection Rate : 0.2102
   Detection Prevalence : 0.2450
      Balanced Accuracy : 0.8103

       'Positive' Class : 1
```





Conclusions

Puzzle, RPG and Indie games got the best reception.

A large portion of games are only supported on Windows.

Top publishers are more likely to have their games succeed.

High sales for a game was determined to be over 20,000 sales.

Linear regression was not a good model for predicting high sales.

Out of the three logistic models, the random forest model has the highest accuracy (89.13%) and AUC (0.94), making it the best-performing model.

References

https://en.wikipedia.org/wiki/List_of_largest_video_game_companies_by_revenue

https://www.kaggle.com/datasets/nikdavis/steam-store-games

Contributions

Kelly Chen: Google Slides creations, EDA, forward selection model

Joshua Petrikat: Regression Models and Analysis, Predictors

Justin Tran: Single factor linear regressions and analysis

Paul Yokota: Rmd file formatting, bug fixes

Bryan Yu: Data sourcing, cleaning, analysis, rmd file formatting