Text Classification of Steam User Reviews

Classifying Reviews Using Steam Data Set

Executive Summary

- Using 9.6 million English reviews from the Steam Library.
- Created a model that can classify user reviews as positive or negative with 91% accuracy.
- The goal of the model is to help classify reviews in order find consumer sentiment and feedback.

Data Features

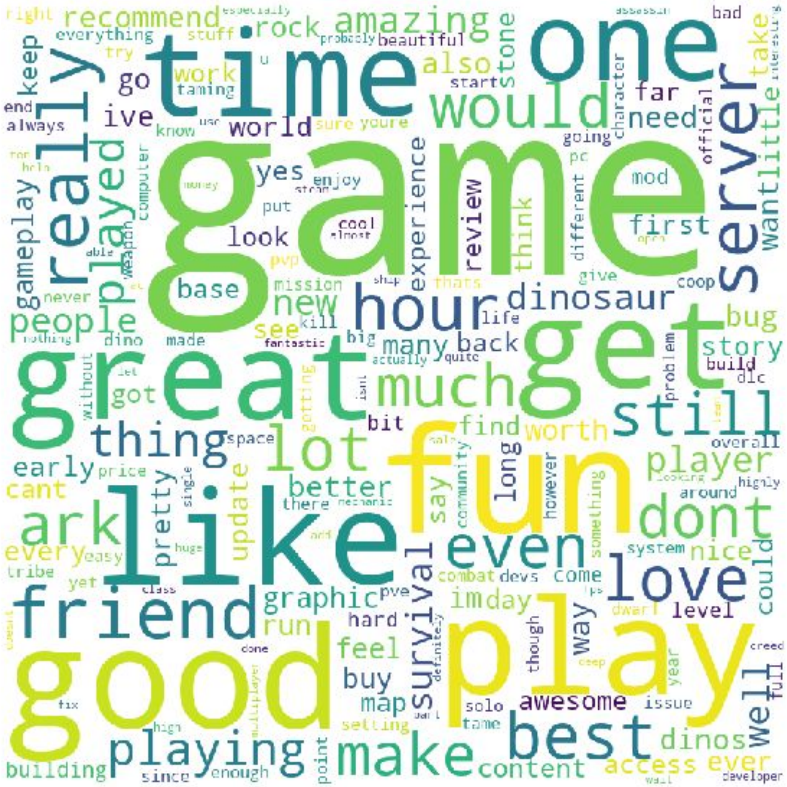
- The Steam dataset consists of reviews in multiple languages, in English there are over 9 million reviews.
- Each review contains different features such as:
  - Game recommended or not
  - Hours user played prior to publishing review
  - Steam purchase
  - Received for free
  - Various user information, such as number of games owned.

Data Wrangling

- User feedback without written reviews were removed.
- A portion of the data was sampled for our model.
  - 3% of games were selected, 9 games total.
  - 263,815 reviews were used.
- Each review was then converted into tokens to separate each review into meaningful units while removing excess information such as stopwords (ex. the, and, then.)

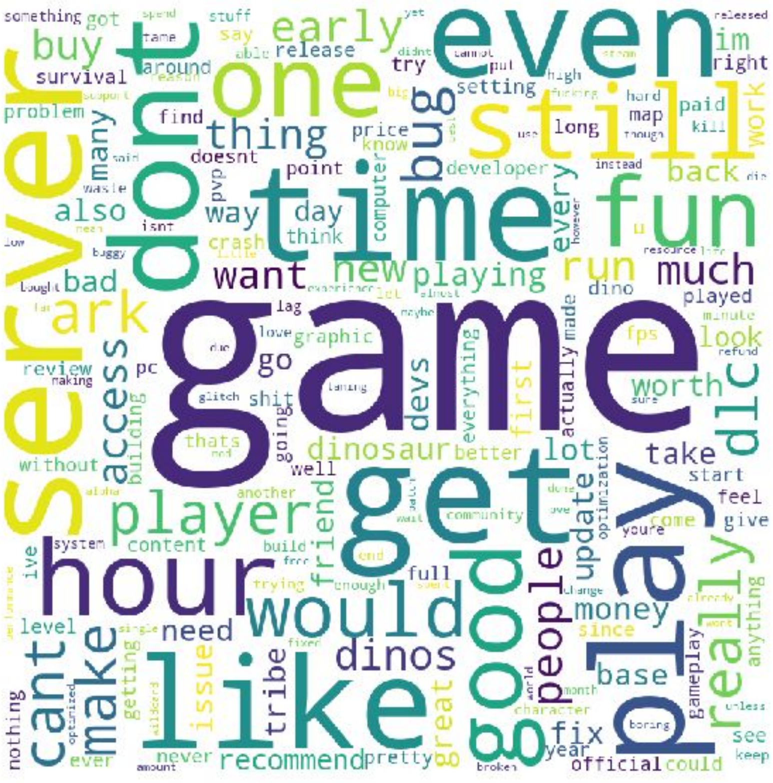
Data Summary-Positive Words

- Positive bigrams and trigrams which have larger effect on our model include:
  - (great, game)
  - (good, game)
  - (best, game, ever)



Data Summary-Negative Words

- Negative bigrams and trigrams which have larger effect on our model include:
  - (early, access)
  - (play, game)
  - (paid, dlc)
  - (early, access, game)



Early Data Analysis - Hypothesis Testing

First Hypothesis: The proportion of games recommended during early access is different than the proportion of games recommended after release date.

Our p-value < α at 1% significance level, two tailed test, we reject the null hypothesis H_0. There is a difference in proportion of positive reviews for early access games and full release games.

		Recommended	
		T	F
Early Access	T	6717	2597
	F	14424	2644

Early Data Analysis - Hypothesis Testing

Second Hypothesis: The proportion of games recommended given for free is different than the proportion of games recommended paid by consumers.

Our p-value < α at 1% significance level, one-tailed test, we reject the null hypothesis H_0. There is a larger proportion of positive reviews for gifted games than proportion of positive reviews for paid games.

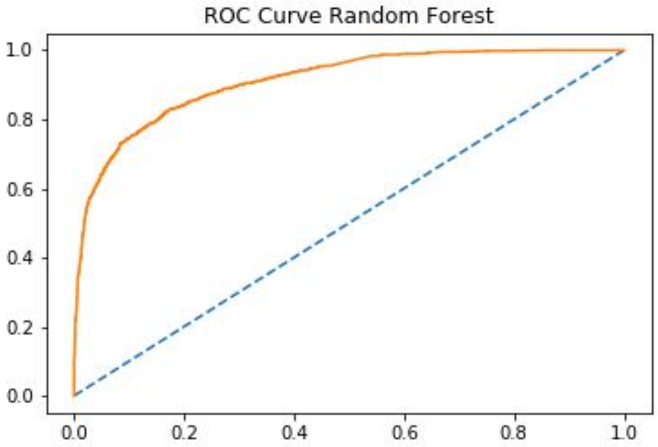
		Recommended	
		T	F
Gifted Game for Review	T	549	99
	F	20592	5142

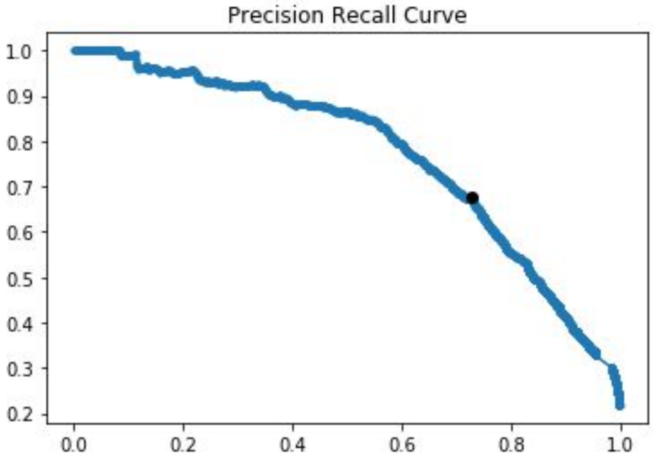
Machine Learning Model

- The Random Forest Classifier performed better than the Multinomial Naive Bayes in a gridsearchCV.
- Using the optimized model to classify the test set resulted an AUC score of 0.911, meaning the model is about 91% correct on its predictions.

Results

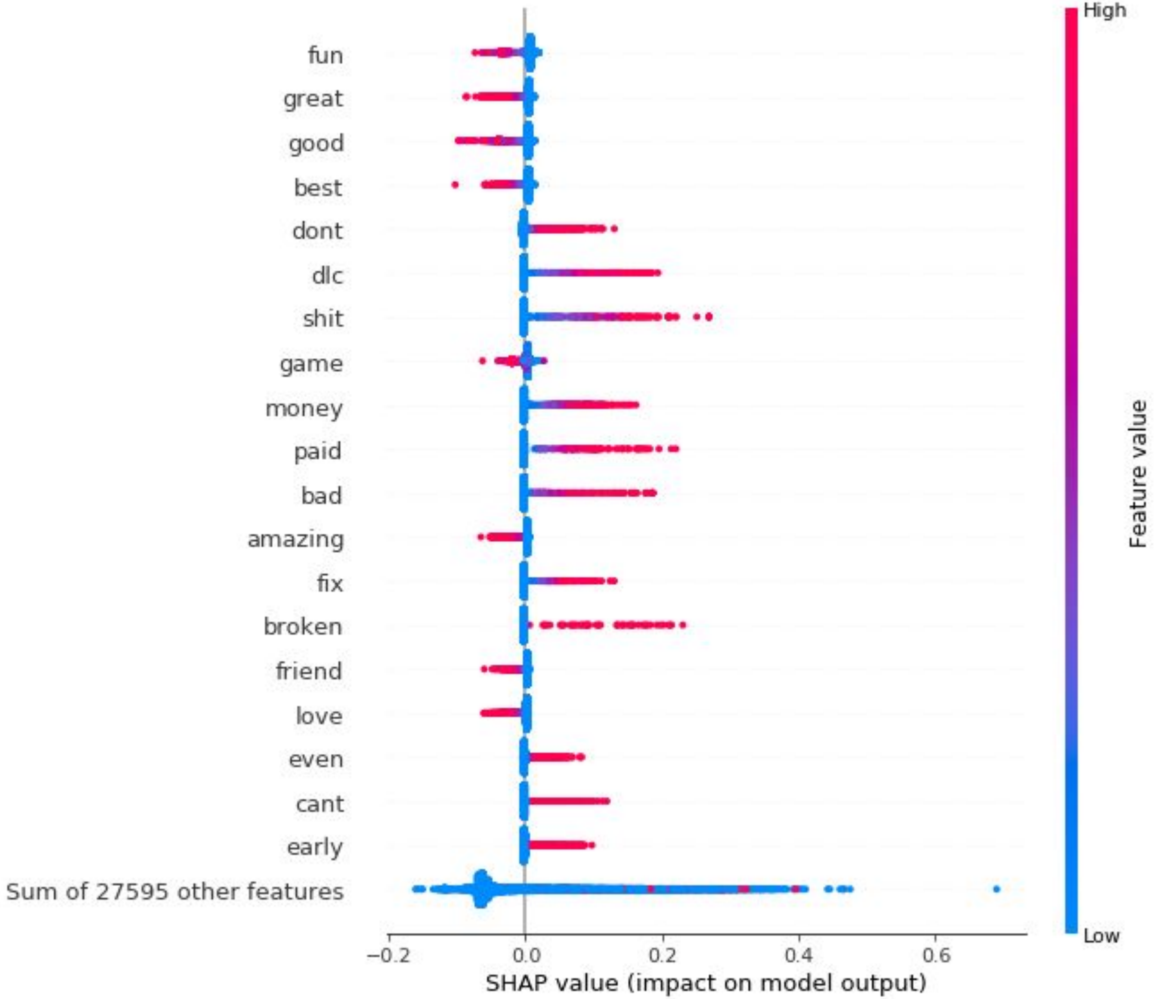
- The ROC curve indicates that the Random Forest performed better than baseline.
- The precision recall curve also depicts that our model has high precision and high recall, but we can also adjust accordingly to our needs, which is the dot on our graph.
- The model is optimized to provide high true positives while minimizing false negatives.





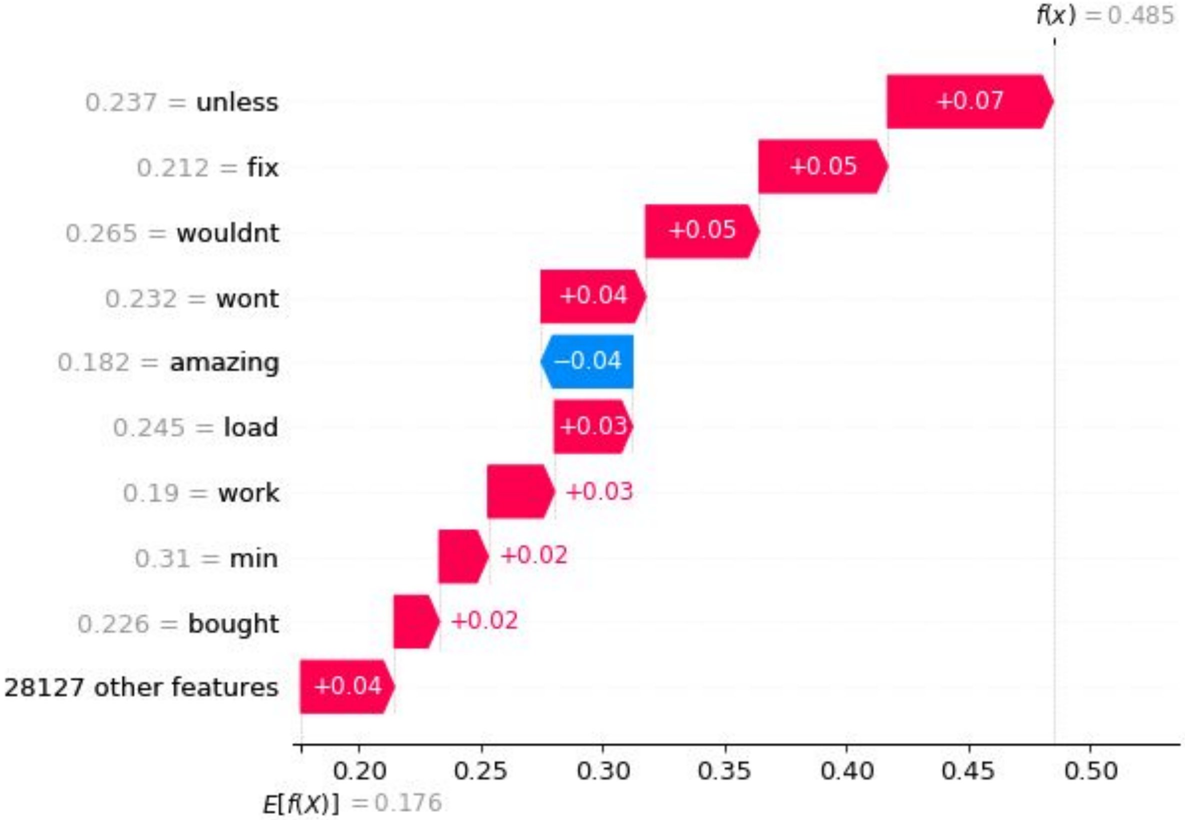
Beeswarm SHAP Plot

- We used SHAP values to visually depict how features affect our model.
- From the table we can see the words in reviews that highly impact the model outputs.
- Most negative reviews contain ‘DLC’, ‘paid’, and ‘broken’. Errors which game developers can look into to adjust or fix.



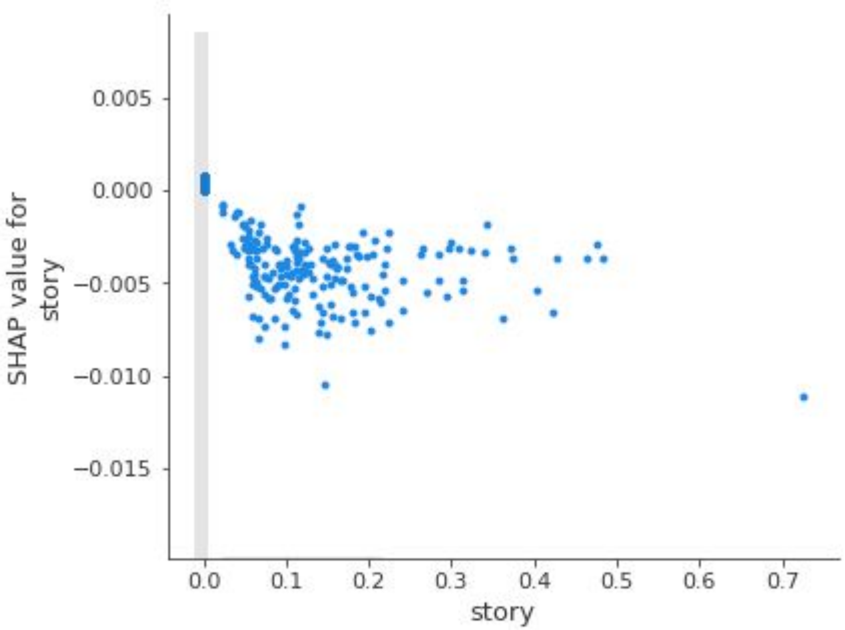
Classification Example

- A consumer wrote the review: “bought game never load computer supasses min spec still wont work seems like amazing concept could ever play wouldnt recommend unless fix problem”
- Our model predicted the review sentiment is negative.
- From our SHAP waterfall plot we can see how each feature contributed to the outcome of our model visually.



Variable Sentiment Search

- Can isolate variables in their own SHAP scatter plots to see their impact on the model.
- For example, if the variable “story” is of importance we can search reviews to see the impact this variable has on our model.
- The scatter plot trends downward meaning each time the variable “story” is present the prediction leans towards a positive sentiment.
- Also worth noting: When the variable “story” is zero or not present, there is a small shift toward a negative review sentiment.



Research Impact

- Quick results without hiring individuals to comb through user reviews and interpret sentiment.
    - Time
    - Human error/bias
    - Accuracy
    - If each review used 3 minutes to process and classify, 20,000 reviews would take 1000 hours at minimum to achieve delaying results, accuracy and efficiency.

Application and Further Research

- Our model can help summarize consumer feedback.
- Quicker consumer feedback results in time saved for bug and programming adjustments.
- The next step in the analysis could include splitting responses into areas of concern. How consumers view these categories could help developers further improve their games.
  - Pricing, DLC, Graphics, etc.