Steam Genre Prediction - Data Science for Business Team Project

Team 44 - Sriya Vemuri, Amna Mahmood, Samir Dar, Kerry Chen, Calida Mathias

TABLE OF CONTENTS

01 Business Understanding

02 Data Understanding

03 Data Preparation and EDA

04 Modeling

05 Evaluation

06 Deployment

01

Business Understanding

About Steam

Steam is the world's ***largest*** digital distribution platform for PC gaming, developed by Valve Corporation. It hosts over ***50,000*** games and serves as the main storefront where developers ***publish and market their games***.

Business Problem

Choosing the right genre tags on Steam store pages affects: Game discoverability and ad targeting; Bundle placement and influencer outreach.

This current process relies on intuition and competitor checks, leading to:

1ST

Missed secondary genres

2ND

Inconsistent tagging across teams

3RD

Delayed Launched

Data Mining Solution

Built a data-mining assistant to predict game genres from early metadata

Inputs:

platform, price, screenshots, trailers, developer history

Outputs:

Ranked genre tags with confidence scores

Offers:

precision vs. reach threshold options

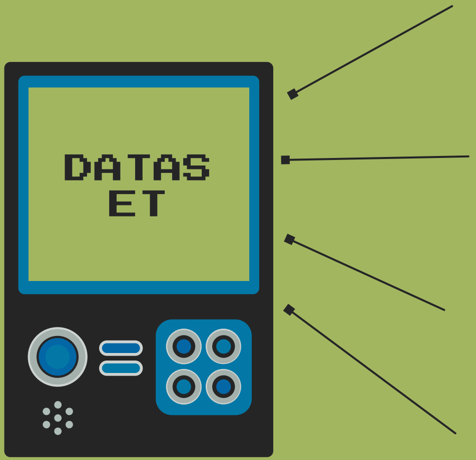
Highlights key drivers:

platform type, price, and media richness

02

Data Understanding

DATA Understanding



DATASET

Source: Merged Steam dataset from Kaggle with ~66,900 games and 31 variables describing title, genre, pricing, platform, and developer details

Data Issues: Originally uncleaned with mixed types, missing values, and semi-structured JSON fields

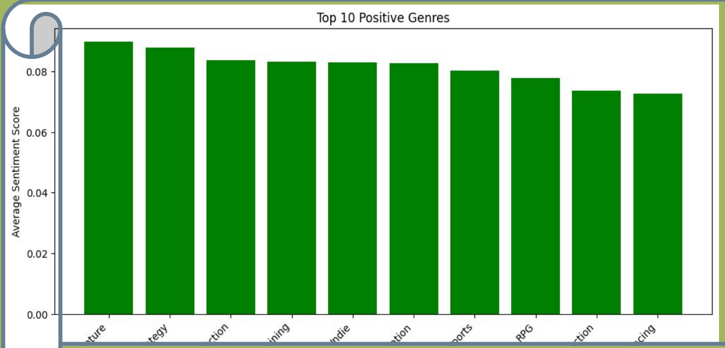
Key Variables: **team_appid, name, genres, price_overview, is_free, required_age, platforms, supported_languages, developers, publishers, screenshots,** and **movies**

Biases & Patterns: Windows games dominate vs. Mac and Linux-based; Genres overrepresented (e.g. Action and Indie); Tagging practices vary across developers (causing class imbalance)

03

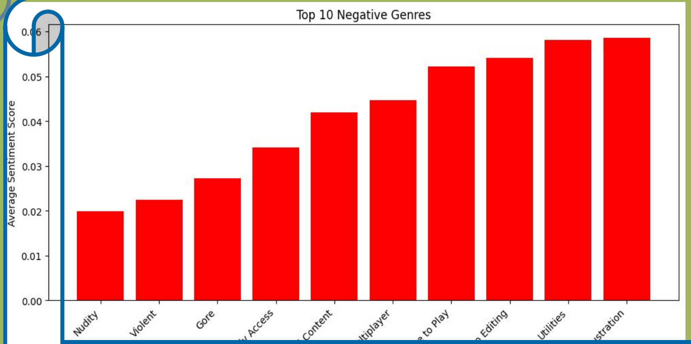
Data Preparation and EDA

Genre Analysis

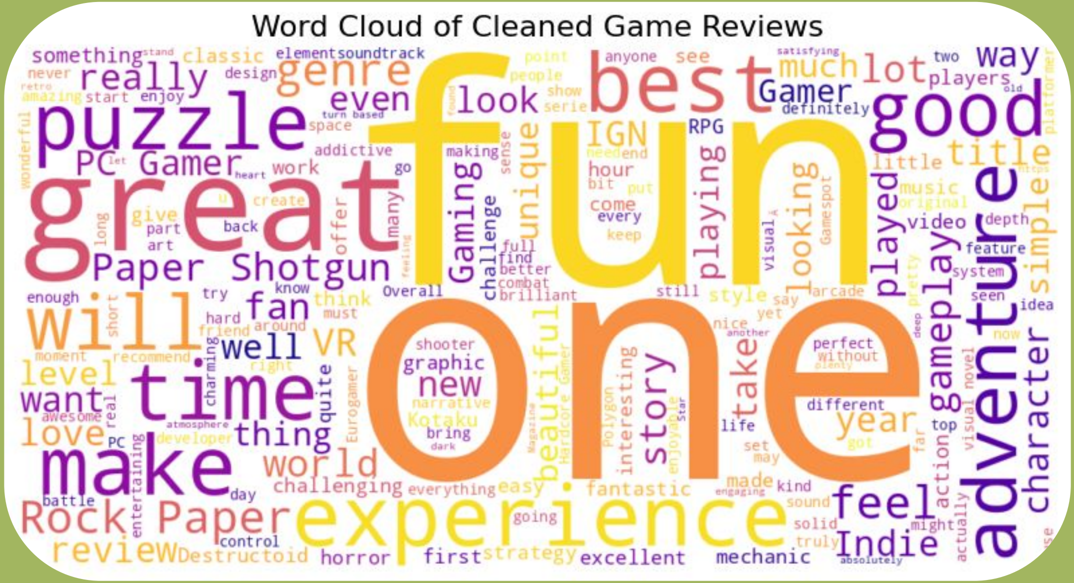


Sentiment reflects engagement differences between gameplay and productivity-focused titles.

Players respond most positively to Adventure and Strategy genres.
Utility and Design apps trend more negatively.



Sentiment Analysis



Word Cloud of Cleaned Game Reviews

04

Modeling

Models Used

Transform early metadata into ranked genre predictions with confidence scores, helping studios make faster, data-driven tagging decisions.

Logistic Regression

Simple, interpretable, shows feature influence

Random Forest

Handles non-linear patterns & missing data well

Linear SVC

Effective for high-dimensional, multi-genre prediction

05

Evaluation

Model Comparison (10-Fold Cross Validation)

~0.5203

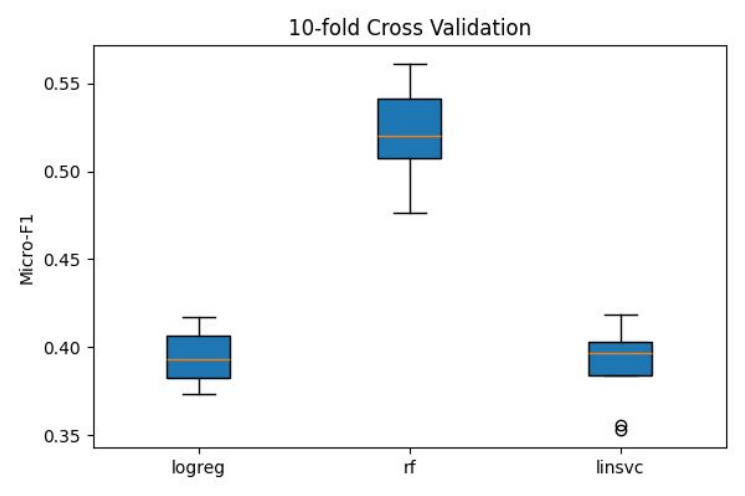
Random Forest

~0.3953

Logistic Regression

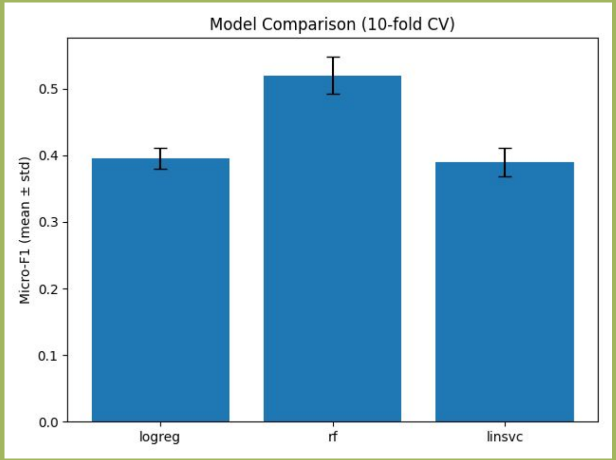
~0.3902

Linear SVC



Metric Used: Micro-F1 score
(mean ± standard deviation)

Interpretation



Performance Summary

Random Forest: Best overall by *Micro-F1* and *Hamming loss*

Fewest per-label tagging errors

Class Imbalance Impact

Macro-F1: low across models (~0.17–0.20)

Rare genres have near-zero recall

Class Imbalance Impact

Weighted-F1: slightly favors **LogReg/SVC**

06

Deployment: A Genre Tagging Assistant

Deployment Plan

- Purpose: help studios and publishers assign the most accurate genres to a new game before launch

developer enters game details

1

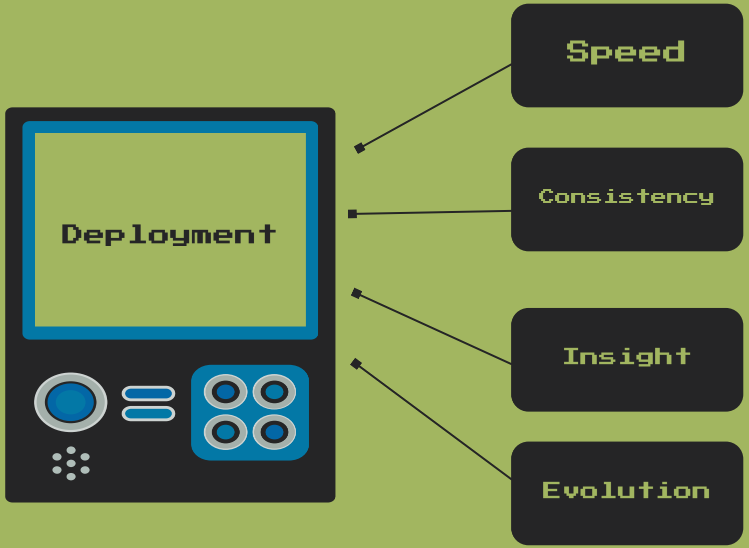
model suggests several likely genres

2

Each suggestion includes a confidence score

3

Benefits



Deployment

Speed: Reduces time spent on manual genre selection

Consistency: Applies a standard, data-driven logic to all titles

Insight: Moves away from guesswork toward analytics

Evolution: Model will be regularly retrained with new Steam data to adapt to market trends and new games

Deployment Considerations

Practical Challenges

Data Quality is Crucial: The model's accuracy depends on complete input. Missing information like pricing or media can weaken its predictions.

Regular Maintenance: The gaming landscape changes fast. The model needs frequent updates to recognize new genres and trends.

User Trust & Transparency: The tool must be easy to use and understand. It should clearly explain *which features* influenced its recommendations to build developer confidence.

Ethical Guidelines

Human in the Loop: This is a support tool, not a final decision-maker. Developers must always review and approve its suggestions.

Bias Awareness: The model learns from existing data and may favor popular genres. Results for niche games must be reviewed with extra care.

Data Integrity: All training data is from publicly available Steam metadata, ensuring no privacy or ethical conflicts.

Risk Mitigation

Outdated Metadata

Implement a schedule for regular model retraining using the latest game data

Bias Toward Common Genres

Balance the training data during updates to give appropriate weight to both popular and niche genres

Over-reliance on Tool

Establish a clear workflow that requires human review and final approval for all genre tags

Technical Complexity

Keep the tool's architecture simple, maintain clear documentation, and ensure it is easy to support

THANK YOU!