RecSys 2015 Purchase Prediction Report

June 2025

Bryan Cheong

MSc Business Analytics

Imperial College London

Part 1: RecSys 2015

Introduction

The report discusses the approach to tackling the RecSys 2015 Challenge, a session-based recommendation task where the goal is to predict whether a user session will result in a purchase, and which items will be bought. The evaluation metric combines session-level classification with item-level accuracy using a custom competition score based on Jaccard similarity and session prediction penalties.

Due to noisy purchase data, class imbalance, and evident temporal drift, we used a hybrid approach:

- XGBoost model for session classification.
- Heuristic model for item prediction (based on recency and popularity).
- Probability threshold and `top_k` tuning to maximise score.

Data Cleaning

The dataset provided consists of `clicks` and `buys` data. Data cleaning was done to ensure consistency and usability of the raw data before later tasks such as feature engineering or modelling. The process is as follows:

1. **Loading and Validating**
   Raw `clicks` and `buys` files were loaded with explicit column names and timestamp parsing. Types were standardised (e.g. converting `Category` to string), and consistency checks ensured that all `buys` sessions had matching `clicks`. Basic dataset statistics were computed to validate shape and coverage.

2. **Price and Quantity**
   Data quality checks on the `buys` dataset revealed that 53% of purchase records had both `Price` and `Quantity` set to zero. These records were deemed corrupted, and all price/quantity-based features were not calculated to prevent training models on unreliable data.

3. **Chronological or Random Split?**
   Two data splitting strategies were initially considered: a realistic *chronological* split to account for the conversion rate decreasing over time, and a *random* split for comparison. However, the project brief explicitly required splitting the data by timestamp, so eventually the chronological split was chosen.

4. **Session-Safe Data Splitting**
   Splits were performed at the session level to prevent data leakage across train, validation, and test sets. Each session was assigned entirely to one split to ensure independence. Sessions were grouped and ordered chronologically by session start time, then allocated based on cumulative record counts to meet approximate ratio targets. Overlap checks confirmed that no session appeared in more than one subset.

5. **Saving Parquet Files and JSON Metadata**
   Cleaned datasets and their session-safe splits were saved in `parquet` format for efficient processing. Metadata for each split (e.g. record/session ratios and time ranges) was stored in a consolidated JSON file to help with reproducibility.

Modelling and Evaluation

Feature Engineering

We engineered a total of 21 features spanning three main categories: session-level behavioural features, item-level statistics, and category-level aggregates. All features were computed using only training data to avoid leakage.

To improve feature consistency and reduce noise, we applied a category cleaning step. Categories with extremely low session counts were grouped under a single placeholder label, RARE_BRAND, while missing or null categories were recoded as UNKNOWN_CAT. The rarity threshold was determined using a data-driven heuristic: we selected the minimum session frequency needed to retain categories that together accounted for 95% of all purchase sessions. This ensured we preserved the most commercially relevant categories while discarding fringe or corrupted entries.

Features such as hour of day and day of week were transformed using sine and cosine functions to capture their natural cyclical patterns. For instance, midnight and 11 PM are represented as close values in this encoding, avoiding artificial discontinuities. Engagement times were also clipped to realistic upper bounds (e.g., one hour) to prevent long-tail outliers from skewing the data.

(a) (b)

Figure 1: Visualisations to show why hour and day of week features are important

More details regarding the features are in the Appendix, the list of features is listed below:

Table 1: Summary of Engineered Features

Feature Category	Feature Name
Session Behaviour	`total_clicks`, `unique_items`, `session_duration`, `engagement_avg/max/std_time`, `start_hour_sin/cos`, `start_dow_sin/cos`
Item Conversion Statistics	`session_item_conversion_rate_mean/min/max`, `session_item_clicks_sum/mean`, `session_item_popularity_score_max/mean`, `session_item_sessions_mean/sum`
Category Aggregates	`session_cat_conversion_rate_mean`, `session_cat_sessions_sum`

Session Classification

We trained an XGBoost classifier with 1000 estimators, which we won't reach due to the addition of early stopping, and scale-adjusted positive class weights to handle severe class imbalance (5.76% positive sessions). The input consisted of the aforementioned engineered features.

Training AUC: `0.7777`
Validation AUC: `0.7049`
Overfitting Gap: `0.0728`

Early stopping typically reduced the number of estimators to fewer than 10, suggesting that only limited boosting was needed.

Threshold and top_k Optimisation

We swept across probability thresholds (0.1 to 0.9, linearly spaced into 20 intervals) and item count parameters (top_k $\in \{1,2,3,5,10\}$) to optimise the custom score:

$$\text{Score} = \sum_{\text{correct sessions}} \left(\frac{|A \cap B|}{|A \cup B|} + \frac{|S_b|}{|S|} \right) - \sum_{\text{wrong sessions}} \frac{|S_b|}{|S|}$$

where A = actual items, B = predicted items, S_b = buy sessions, S = all sessions.

Optimal Probability Threshold: `0.563`
Best top_k: `10`
Validation Score: `3407`

Item Prediction Heuristics

Instead of training a model to predict items, we used a deterministic method:

- For each clicked item: score = recency weight $\times$ item conversion rate.
- Select top_k items per session based on this score.

This approach generalised better than learning-based methods, which tend to overfit.

Visual Analysis

Threshold Effects

Figure 2 presents several diagnostic plots that were used to guide the selection of the optimal decision threshold.

- **Top-Left:** Illustrates how the custom competition score varies across different probability thresholds. The peak marks the best trade-off between true positives and penalties for incorrect predictions.
- **Top-Right:** Shows how many sessions were classified as purchases under each threshold. As expected, higher thresholds lead to fewer positive predictions.
- **Bottom-Left:** Visualises the trade-off between precision and recall. Our selected threshold lies on a point that moderately balances the two, with relatively low precision but acceptable recall.
- **Bottom-Right:** Displays the distribution of predicted probabilities for purchase (red) and non-purchase (blue) sessions. The dashed green line marks the optimal classification threshold at 0.563, which was selected by maximising the custom competition score. This value effectively balances false positives and false negatives by isolating a denser region of purchase sessions (right of the line) while filtering out most of the non-purchase noise (left of the line).

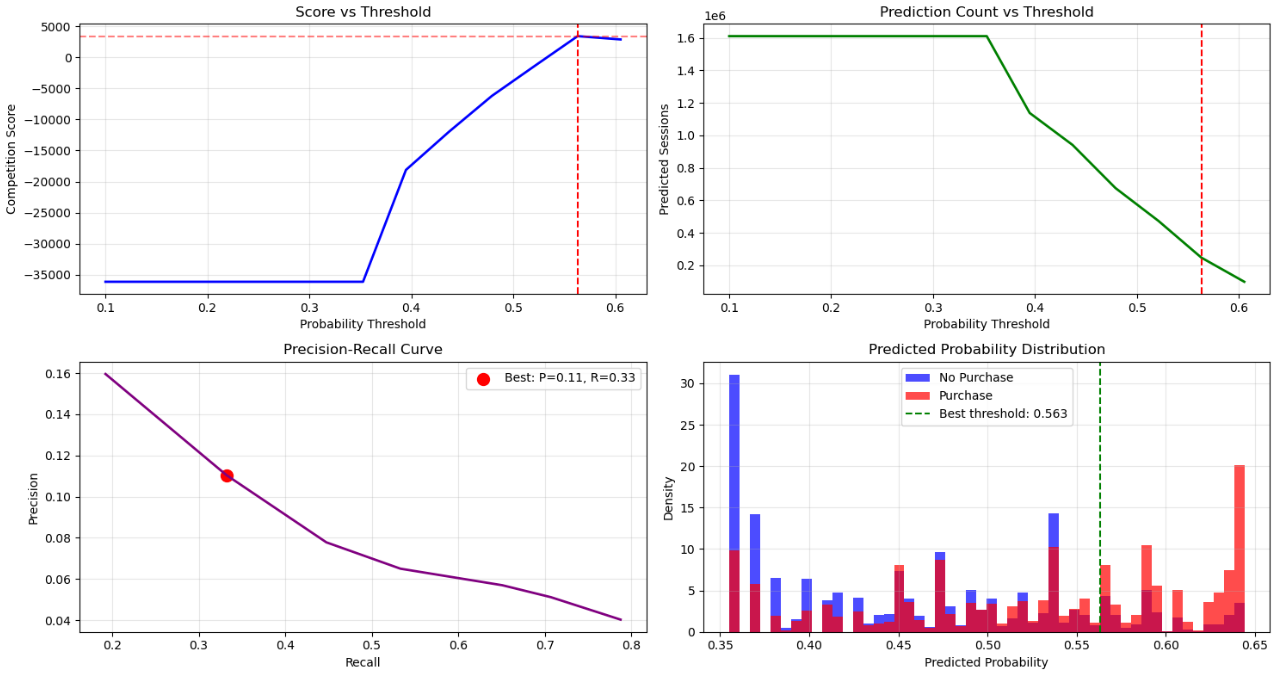

Figure 2: Visual diagnostic plots for threshold tuning.

SHAP Analysis

To interpret the session classifier, we used SHAP (SHapley Additive exPlanations) values computed on the validation set. The summary plot in Figure 3 shows the top features ranked by their average absolute impact on model output.

The most influential factors include:

- `session_duration`: longer sessions were more indicative of purchase intent;
- `session_item_conversion_rate_mean`: sessions with a history of higher item conversion rates tended to result in purchases;
- `total_clicks`: overall engagement levels proved important.

We observed that only a subset of features made substantial contributions to the model's predictions, while many others had a negligible impact.

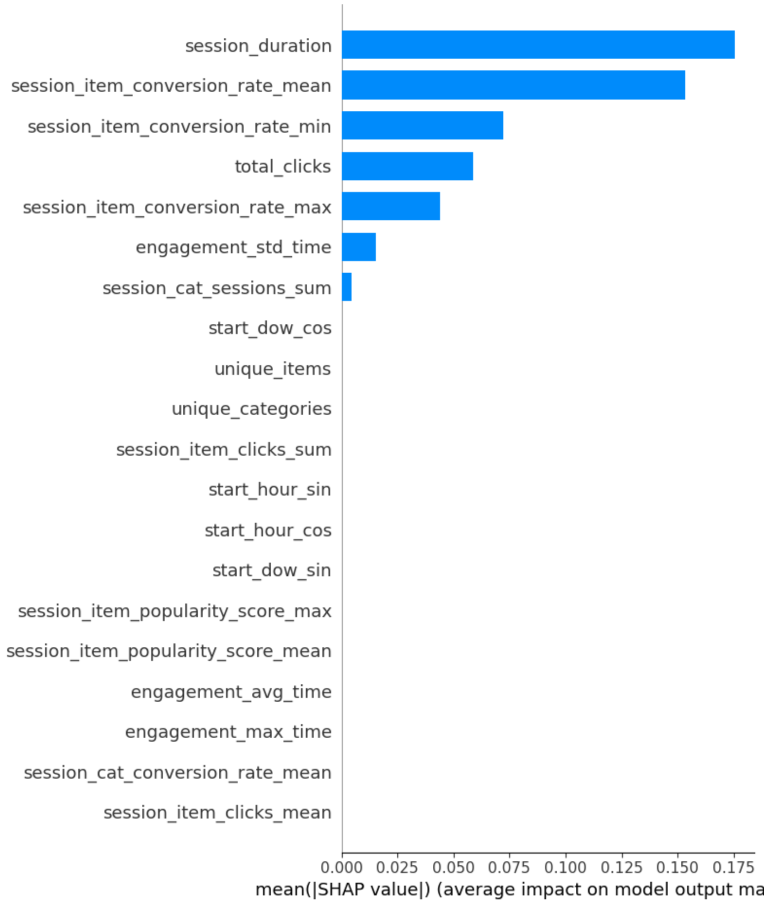

Figure 3: SHAP Summary Plot for XGBoost Session Classifier

Final Score

The final evaluation on the held-out test set, using the competition-specific scoring metric, resulted in a score of **10491.3**.

Conclusion

We developed a pipeline for predicting purchase sessions and associated items in the presence of strong label imbalance and significant noise. While various modelling approaches were considered for both stages, heuristics proved more robust than learning-based methods for item prediction. The final solution combines an XGBoost-based session classifier with a deterministic, score-weighted heuristic for item selection. This ensemble delivered a competitive performance using the custom evaluation metric while maintaining simplicity, interpretability, and generalisability.

Part 2: Amazon FTC Case Study

Introduction

This essay examines the case of Federal Trade Commission et al v. Amazon.com (2023), in which the Federal Trade Commission (FTC) and 17 U.S. states brought a legal action against Amazon (Federal Trade Commission 2023). It was argued that Amazon unlawfully maintains monopoly power in two markets: the online superstore market and the online marketplace services market, through a series of illegal and anticompetitive practices. This essay argues that the FTC and the 17 U.S. states are likely to succeed in their case against Amazon. The claim is supported by the facts that Amazon's conduct does not merely reflect market dominance but rather a deliberate strategy aimed at suppressing competition and entrenching its monopoly power (Khan 2017). By examining the allegations against Amazon, this essay will demonstrate how Amazon's practices contravene established principles and why regulatory intervention is warranted.

Facts

Amazon engages in a practice known as 'anti-discounting,' which allows it to monitor and detect when sellers offer the same product at a lower price on other online platforms. As a response, Amazon punishes these sellers to prevent competitors from attracting customers with lower prices. These punishments are not limited to lowering their prices to compete with them, but also include suppressing their visibility in search results, removing price information, or excluding them from the critical 'Buy Box,' which significantly affects their sales.

Additionally, Amazon forces its sellers to use its Fulfilment by Amazon (FBA) service to qualify for Prime eligibility, preventing sellers from choosing independent logistics providers other than their own and limiting competition in the fulfilment market. Amazon also manipulates pricing through an algorithm codenamed 'Project Nessie,' which raises prices on select products, prompting competitors to match and generating over $1 billion in profit (Federal Trade Commission 2023). These practices have degraded the user experience by flooding search results with paid advertisements, which prioritise sponsored listings at the expense of relevance (Federal Trade Commission 2021). These practices have collectively entrenched Amazon's market dominance, inflated consumer prices, suppressed competition, and stifled innovation. This case is not solely about Amazon's dominance, but about the unlawful methods they used to obtain and exploit that power, raising concerns about the balance between innovation, fair competition, and market control in the digital economy.

Legal Aspects

Before examining the relevant laws and regulations, it is important to consider legal precedent to understand how courts distinguish between legitimate competitive conduct and unlawful monopolistic behaviour. In the case of United States of America v. Microsoft Corporation (2001), Microsoft was found to have maintained its monopoly by restricting the abilities of manufacturers and users to remove Internet Explorer or use alternative browsers (*United States v. Microsoft Corporation* 2001). These barriers limited the consumer choices and prevented competitors from challenging their dominance. The court held that while holding monopoly power is not illegal per se, using that monopoly power to suppress competition through anti-competitive practices constitutes a violation of Section 2 of the Sherman Act, which prohibits monopolisation and attempts to monopolise in the United States (Hovenkamp 2021). Amazon's practices closely mirror those of Microsoft, where it uses its dominance to suppress price competition, forcing sellers to use its fulfilment services or manipulate pricing in ways that disadvantage their competitors. Given the similarity of the two cases and the growing legal scrutiny of digital monopolies, it is likely that the courts will hold Amazon's conduct as unlawful and harmful to market competition.

A notable example is Amazon's 2010 acquisition of Quidsi, the parent company of Diapers.com (Stone 2013). After Quidsi initially rejected Amazon's acquisition offer, Amazon retaliated by aggressively cutting prices on baby products. It launched the 'Amazon Mom' program, offering free shipping and steep discounts, which resulted in significant losses in the diaper category within a few months. This predatory pricing strategy slowed Quidsi's growth and ultimately pressured the company to accept Amazon's offer out of fear of market suppression (Khan 2017). Following the acquisition, Amazon significantly reduced the discounts previously offered through Amazon Mom. This example has clearly illustrated how Amazon uses its market dominance to eliminate a competitor through 'below-cost pricing' and withdraws the benefits once it has acquired the company or eliminated the competitor. Such practices align with what is known as 'predatory pricing', a form of exclusionary conduct that violates Section 2 of the Sherman Antitrust Act. This strategy reveals their intention to employ anti-competitive tactics to eliminate competitors from the market, which supports the concerns raised by the FTC and 17 states that Amazon's dominance is maintained not through innovation, but through coercive and anti-competitive means.

In addition to violating Section 2 of the Sherman Act, Amazon's conduct falls within the scope of Section 5 of the Federal Trade Commission (FTC) Act, which prohibits unfair practices in competition. The FTC argued that Amazon's strategy, including punishing sellers for offering lower prices on other platforms, forcing sellers into using Fulfilment by Amazon to qualify for Prime, and manipulating search rankings to favour paid listings, constitutes exclusionary conduct that distorts market transparency and undermines fair competition. These practices harm not only their competitors but also their consumers, who are left with fewer choices and higher prices. These practices have collectively undermined fair competition and reinforced their market dominance, making a strong case for regulatory intervention.

Digital Marketing Analytics Insights

Amazon strategically employs advanced digital marketing analytics to consolidate its market dominance through sophisticated targeting, personalisation, and competitive intelligence strategies. Leveraging extensive first-party data, Amazon integrates precise consumer insights into powerful advertising platforms like Amazon DSP and the Amazon Marketing Cloud that provide significant competitive advantages over traditional retailers (Mattioli 2019). For instance, Amazon DSP and Amazon Marketing Cloud have enabled advertisers, such as a large U.S. telecom brand, to achieve approximately a 40% reduction in cost-per-action (CPA) by leveraging their first-party signals for highly targeted campaigns (Advertising 2024).

Utilising analytics frameworks such as RFM (Recency, Frequency, Monetary) segmentation, Amazon identifies and prioritises high-value customers, tailoring personalised promotional content to maximise engagement and sales conversion rates. Amazon's Project Nessie algorithm exemplifies how sophisticated pricing analytics can create tacitly collusive environments, using predictive modelling to orchestrate competitor price increases (Singh et al. 2023). Notably, Project Nessie's pricing tactics reportedly generated over $1 billion in additional profits (Federal Trade Commission 2023).

A 2011 pricing algorithm incident, where two booksellers' rule-based systems inflated a book's price to $23.7 million, demonstrates how even simple, rule-driven algorithms can trigger runaway price escalations. By covertly coordinating pricing outcomes, Amazon's system crosses the line from automated efficiency into anticompetitive strategy (Solon 2011).

Amazon also benefits from strong direct and indirect network effects (Rochet & Tirole 2003). More customers attract more sellers, and more sellers improve variety and convenience, further attracting customers. This self-reinforcing loop creates a powerful platform moat that deters entry by rivals (Parker et al. 2016). From a digital marketing perspective, this magnifies Amazon's customer lifetime value (CLV) and grants it unmatched cost-per-acquisition (CPA) efficiency, making it nearly impossible for smaller sellers to compete effectively in advertising ROI. The FTC could argue that Amazon isn't just reaping the benefits of scale; it is artificially locking in these advantages through data dominance and platform constraints, such as requiring Fulfilment by Amazon (FBA) for Prime eligibility.

Furthermore, Amazon's growing retail media network has transformed product discovery into a payto-play environment. Sponsored listings dominate the results page, crowding out organic results and undermining relevance-based UX (Amaldoss 2025). From a digital marketing analytics standpoint, this shifts the discovery model from one based on user intent (click-through rates, relevance) to an auction-based visibility system. This not only skews consumer experience but also marginalises sellers who lack the advertising budget, reinforcing Amazon's dominance.

Amazon's extensive data harvesting also enables exclusionary feedback loops. By analysing seller performance, pricing patterns, demand elasticity, and customer behaviour, Amazon can identify high-performing third-party products and then introduce Amazon-branded alternatives (e.g., Amazon Basics) with preferential placement (Reuters 2021). This dual role as both platform operator and market participant blurs competitive lines. The use of digital analytics not to improve consumer outcomes, but to eliminate competitors through predictive advantage, exemplifies competitive foreclosure, a concern echoed by regulators in India and the European Union.

These analytics-driven practices significantly distort competitive conditions, elevate market entry barriers, and contribute to degraded consumer experiences by prioritising sponsored content over organic search results. Consequently, these practices have sparked substantial regulatory concerns regarding transparency, competitive fairness, and the ethical use of analytics in digital marketplaces.

Amazon's Counter-Arguments and Defensive Position

Amazon's defense strategy centers on demonstrating tangible consumer benefits from its platform operations (Evans & Schmalensee 2016). The company argues that its integrated model delivers unprecedented convenience, selection, and pricing advantages through operational efficiencies rather than anticompetitive behaviour. Amazon maintains that dynamic pricing algorithms serve legitimate business purposes by optimising inventory management and responding to market conditions (Brynjolfsson et al. 2019).

Amazon challenges market definition disputes, arguing that the relevant market extends beyond online retail to include all retail channels (Katz & Sallet 2018). Under this broader definition, Amazon's market share appears smaller, potentially undermining monopolisation claims. The company points to competition from Walmart, Target, and emegring platforms as evidence of market contestability.

However, these arguments face scrutiny given the technical evidence of algorithmic manipulation and the documented pattern of exclusionary practices that extend beyond normal competitive responses (Federal Trade Commission 2023).

Business Impact Analysis

The case represents fundamental tensions in how digital platforms reshape entire industries. Amazon's platform hosts over 2 million active sellers, creating unprecedented conflicts of interest where the platform simultaneously serves as an infrastructure provider, competitor, and rule-maker (Khan 2017). Small and medium-sized sellers experience "platform dependency," with businesses becoming increasingly reliant on Amazon's traffic and logistics.

The business impact manifests through margin compression via increasing advertising costs and fulfilment fees, effectively transferring wealth from merchants to the platform. Amazon's advertising revenue grew from \$1.7 billion in 2016 to over \$31 billion in 2022, largely extracted from sellers (Amazon.com Inc. 2022). Sellers also face "innovation capture", where Amazon uses seller data to launch competing private-label alternatives with preferential placement (Mattioli 2019).

Critical Examination of Proposed Remedies

Proposed remedies to Amazon's conduct, including algorithm transparency, platform-seller separation, and third-party logistics rights, offer potential pathways to restore competition, but they pose complex implementation challenges.

Algorithm transparency could enhance oversight of price manipulation and search ranking bias. However, as noted by Pasquale (2015) and Burrell (2016), modern algorithms are often opaque and dynamic, complicating external audit efforts. Moreover, forced disclosure risks undermining competitive confidentiality or enabling tacit collusion by competitors.

Platform-seller separation addresses structural conflicts of interest but assumes Amazon's roles can be cleanly disentangled. In practice, this may reduce product variety or drive up costs for consumers if Amazon scales back private-label offerings or raises seller fees (Crémer et al. 2019).

Independent logistics remedies could foster competition in fulfilment but may weaken Prime's reliability if quality standards differ. Crafting neutral eligibility rules that preserve service consistency while supporting seller choice remains a significant challenge (Schweitzer et al. 2018).

Audit requirements for advertising and search fairness offer promise but are technically demanding. Fairness is subjective, and audits risk being gamed or generating compliance theatre unless tied to measurable outcomes.

Alternative strategies such as behavioural rules, interoperability mandates, or data portability requirements may offer more flexible tools for addressing market power without compromising innovation. In addition, international regulatory coordination will be essential to avoid simply shifting market power to less-regulated environments.

Case Prediction

Given the strength and breadth of the evidence presented by the FTC, it is likely that regulators will succeed in their case against Amazon. The allegations are not merely speculative but are supported by specific examples of conduct that reflect deliberate and systemic anticompetitive strategies. The documented practices point to systematic exclusionary behaviour designed to restrict competition. These actions violate both the Sherman Act and the FTC Act by maintaining market power through coercive rather than competitive means (Wu 2018).

What strengthens the FTC's case further is the digital traceability and algorithmic evidence that supports the regulators' claims. Amazon's internal documentation, algorithm deployment, and data-driven enforcement mechanisms offer concrete insights into intent and market manipulation. In an era where digital marketing and platform dynamics are increasingly central to commerce, the regulators are also likely to benefit from a broader policy context that favours increased scrutiny of dominant tech firms (Newman et al. 2024).

However, Amazon's counter-arguments regarding consumer welfare and market definition present significant challenges to the FTC's case. The company's ability to demonstrate concrete consumer benefits and point to competitive pressures from alternative retail channels may resonate with courts traditionally focused on consumer welfare standards. The complexity of digital markets and the technical nature of the evidence may also create challenges for regulatory success, particularly if Amazon can effectively argue that its practices reflect operational necessities rather than anticompetitive intent.

Conclusion

The case against Amazon reflects a critical inflection point in how digital markets are regulated. While innovation and efficiency are hallmarks of successful digital enterprises, the unlawful means by which Amazon has allegedly maintained its dominance challenge the core principals of market fairness and consumer welfare. Through coercive tactics like anti-discounting, forced fulfilment services, algorithmic price manipulation, and data-driven exclusionary practices, Amazon has strategically entrenched its monopoly power. These actions not only harm competition but also degrade the consumer experience and restrict innovation. However, effective remedies must address the complexity of digital platform regulation while preserving innovation incentives.

The case ultimately represents a broader challenge for competition policy in digital markets: how to preserve the benefits of platform innovation while preventing the abuse of market power. Success will require nuanced approaches that address specific anticompetittive behaviours without undermining the fundamental efficiencies that make digital platforms valuable to consumers and market participants. Given the compelling evidence of Sherman Act and FTC Act violations, regulatory intervention apppears justified, though remedies must be carefully crafted to preserve innovation while restoring competition.

References

Advertising, A. (2024), 'When a large telecom brand leveraged first-party signals to achieve a 40 Accessed 29 June 2025.
URL: *https://advertising.amazon.com/library/case-studies/amazon-marketing-cloud-telcom-iprospect*

Amaldoss, W. (2025), 'Why amazon's private labels hurt consumers more than third-party sellers', *Duke Fuqua Insights* .
URL: *https://www.fuqua.duke.edu/duke-fuqua-insights/Amazon-private-labels-hurt-consumers*

Amazon.com Inc. (2022), Annual report 2022, Technical report, Amazon.com Inc., Seattle.

Brynjolfsson, E., Hu, Y. J. & Rahman, M. S. (2019), *Competing in the age of omnichannel retailing*, MIT Sloan Management Review.

Burrell, J. (2016), 'How the machine 'thinks': Understanding opacity in machine learning algorithms', *Big Data & Society* **3**(1), 1–12.

Crémer, J., de Montjoye, Y.-A. & Schweitzer, H. (2019), Competition policy for the digital era, Technical report, European Commission.

Evans, D. S. & Schmalensee, R. (2016), *Matchmakers: The new economics of multisided platforms*, Harvard Business Review Press.

Federal Trade Commission (2021), 'Amazon search results are full of ads 'unlawfully deceiving' millions of consumers, ftc complaint says', The Washington Post.
URL: *https://www.washingtonpost.com/technology/2021/12/08/amazon-search-results-ftc-complaint/*

Federal Trade Commission (2023), 'Federal trade commission et al. v. amazon.com, inc.', FTC Complaint.

Hovenkamp, H. (2021), *Federal antitrust policy: The law of competition and its practice*, 6 edn, West Academic Publishing.

Katz, M. L. & Sallet, J. (2018), 'Multisided markets and antitrust enforcement', *Yale Law Journal* **127**(7), 2142–2175.

Khan, L. M. (2017), 'Amazon's antitrust paradox', *Yale Law Journal* **126**(3), 710–805.

Mattioli, D. (2019), 'Amazon scooped up data from its own sellers to launch competing products', *The Wall Street Journal* .

Newman, J. M. et al. (2024), 'Understanding the tradeoffs of the amazon antitrust case', *Harvard Business Review* .
URL: *https://hbr.org/2024/01/understanding-the-tradeoffs-of-the-amazon-antitrust-case*

Parker, G. G., Van Alstyne, M. W. & Choudary, S. P. (2016), *Platform revolution: How networked markets are transforming the economy and how to make them work for you*, W. W. Norton & Company.

Pasquale, F. (2015), *The black box society: The secret algorithms that control money and information*, Harvard University Press.

Reuters (2021), 'Amazon copied products and rigged search results to promote its own brands', *Reuters* .

Rochet, J.-C. & Tirole, J. (2003), 'Platform competition in two-sided markets', *Journal of the European Economic Association* **1**(4), 990–1029.

Schweitzer, H., Haucap, J., Kerber, W. & Welker, R. (2018), Modernising the law on abuse of market dominance, Technical report, German Federal Ministry for Economic Affairs and Energy.

Singh, P. V., Wang, Q., Huang, Y. & Srinivasan, K. (2023), 'Algorithms, artificial intelligence and simple rule-based pricing', Working paper via Carnegie Mellon University. Studied algorithms for pricing and their tacit collusion behavior.
URL: *https://www.cmu.edu/news/stories/archives/2023/october/algorithmic-pricing-understanding-the-ftcs-case-against-amazon*

Solon, O. (2011), 'How a book about flies came to be priced $24 million on amazon', *Wired* .
URL: *https://www.wired.com/2011/04/amazon-flies-24-million*

Stone, B. (2013), *The everything store: Jeff Bezos and the age of Amazon*, Little, Brown and Company.

United States v. Microsoft Corporation (2001), 253 F.3d 34 (D.C. Cir.).

Wu, T. (2018), *The curse of bigness: Antitrust in the new gilded age*, Columbia Global Reports.

Appendix

Table 2: Feature Descriptions for Session Classification

Feature Name	**Description**
total_clicks	Total number of clicks in the session.
unique_items	Number of distinct items clicked in the session.
unique_categories	Number of distinct categories encountered in the session.
session_duration	Time span (in seconds) between the first and last click.
engagement_total_time	Sum of time spent between clicks. Capped at 1 hour per click.
engagement_avg_time	Mean engagement time across all clicks in the session.
engagement_max_time	Maximum engagement time between any two consecutive clicks.
engagement_std_time	Standard deviation of engagement times.
start_hour_sin/cos	Cyclic encoding of session start hour (to capture daily patterns).
start_dow_sin/cos	Cyclic encoding of session start day of week.
session_item_conversion_rate_mean	Mean item conversion rate across all clicked items.
session_item_conversion_rate_max	Maximum item conversion rate in the session.
session_item_conversion_rate_min	Minimum item conversion rate in the session.
session_item_clicks_sum	Total global clicks for all items in the session.
session_item_clicks_mean	Average number of clicks per item across all sessions.
session_item_popularity_score_max	Maximum item popularity score (rank percentile).
session_item_popularity_score_mean	Average popularity score of clicked items.
session_item_sessions_sum	Sum of unique sessions for all clicked items.
session_item_sessions_mean	Mean number of sessions in which clicked items appeared.
session_cat_conversion_rate_mean	Mean conversion rate of encountered categories.
session_cat_sessions_sum	Total number of sessions for all encountered categories.
label	Target variable: 1 if purchase occurred, else 0.
Session_ID	Unique session identifier (used for grouping and scoring).