

GROUP 16

**KOL Selection to Maximize Campaign GMV
Using Data-Driven Modeling and Optimization**

Under the guidance of
Dr. Ehsan Ahmadi
Stetson-Hatcher School of Business
Mercer University

By:
San Nguyen - 11052408
Thanh Vo - 11056785
Neha Chowke - 11056797

Table of contents

1. Introduction	1
1.1 Executive Summary	1
2. Problem Description	1
2.1. The Rise of Shoppertainment and Livestream Commerce	1
2.3. Challenges in KOL Selection and Campaign Optimization	2
2.4. A Data-Driven Solution for ROI Optimization	2
3. Background	2
4. Literature Review	3
5. Exploratory Data Analysis (EDA)	3
6. Model Formation	5
6.1. Linear regression	5
6.2. Linear regression with logarithmic transformation	6
6.3. Optimization Model: Business Scenario and Mathematical Formulation	7
6.4. Model Flexibility and Adaptability	9
7. Business Impact	9
7.1. Evaluation of Manual Selection Versus Model-Based Selection	9
7.2. Strategic Implications	10
7.3. Recommendations	10
8. Theoretical Frameworks and Methodological Strategies	11
8.1. Predictive Modeling in KOLs:	11
8.2. Algorithmic Customization and Social Influence	11
9. Strategic Recommendations for Campaign Optimization	11
10. Conclusion and Recommendation	12
6. Appendix	13

1. Introduction

Exploring a data-driven approach to optimizing Key Opinion Leader (KOL) selection for TikTok influencer marketing campaigns. With the rapid growth of "shoppertainment" — a blend of shopping and entertainment — brands face increasing pressure to allocate budgets efficiently while maximizing engagement and Gross Merchandise Value (GMV). Our objective is to replace manual, intuition-based KOL selection with a scalable, automated framework powered by statistical modeling and optimization. By leveraging linear regression and prescriptive analytics, the project identifies the most impactful KOLs under specific budget and engagement constraints, ultimately improving campaign ROI and setting a foundation for smarter future campaigns.

1.1 Executive Summary

This project solves the problems related to TikTokShop KOL marketing campaign performance by implementing a predictive and prescriptive optimization model in place of a KOL manual selection process. The current model KOL selection is rather intuitive and based on prior interactions, which leads to poor ROI and inconsistent engagement. From the provided dataset of 2000+ TikTok influencers, a linear regression model was built to predict GMV performance. Subsequently, a constraint-based optimization model was developed for KOL selection and content activity allocation.

When testing both approaches using the same budget of \$7,500, the optimized model surpassed the manual model in predicting GMV by 221% (from \$114,849.53 to \$368,367.98). While the manual model tends to underperform, the optimized model performs better and continues to meet engagement targets, cost ratios, and KOL tier diversity. Our model provides KOL adaptable solutions that can endure the test of time by integrating new data on cost and engagement, ensuring enduring marketing results in the age of shoppertainment.

2. Problem Description

2.1. The Rise of Shoppertainment and Livestream Commerce

The integration of shopping and entertainment through livestreams and short-form videos, known as shoppertainment, has transformed digital commerce across Asia. TikTok Shop and Shopee KOLs have been turned into real-time sales agents who promote products via engaging content and make commission only if sales are made. Although this affiliate model shifts a merchant's financial risk by minimizing upfront costs, it creates other challenges in campaign planning and selection of KOLs.

2.3. Challenges in KOL Selection and Campaign Optimization

The existing approach to KOL selection is largely manual, relying on social reach or subjective past collaborations. With over 2,000 potential KOLs in a typical campaign pool, this process becomes inefficient and inconsistent. As a result, many campaigns face: poor return on investment (ROI) due to non-strategic budget allocation, missed Gross Merchandise Volume (GMV) targets, unpredictable engagement outcomes. This selection process is often likened to "searching for a needle in a haystack"—a time-intensive and non-scalable task

2.4. A Data-Driven Solution for ROI Optimization

To resolve these inefficiencies, our project proposes a predictive and prescriptive framework that:

- Predicts KOL performance using variables such as booking cost, follower base, and engagement rate.
- Automates selection and content allocation through optimization modeling.
- Incorporates business constraints like budget limits, minimum engagement thresholds, and tier diversity.
- By adopting such a standardized approach, organizations can more effectively tier influencers, make transparent allocation decisions, and benchmark across teams or competitors using shared criteria

3. Background

With the rise of *shoppertainment*—a fusion of shopping, entertainment, and social media—platforms like TikTok have transformed how consumers engage with products online. In this new commerce landscape, Key Opinion Leaders (KOLs) play a critical role in influencing purchasing decisions through livestreams, short videos, and authentic product endorsements. However, businesses currently rely on manual and inconsistent methods to select KOLs, often leading to suboptimal budget use and unpredictable campaign outcomes. This project was initiated to address that inefficiency by building a data-backed, scalable solution for selecting and allocating KOLs to maximize Gross Merchandise Value (GMV) while meeting specific campaign goals.

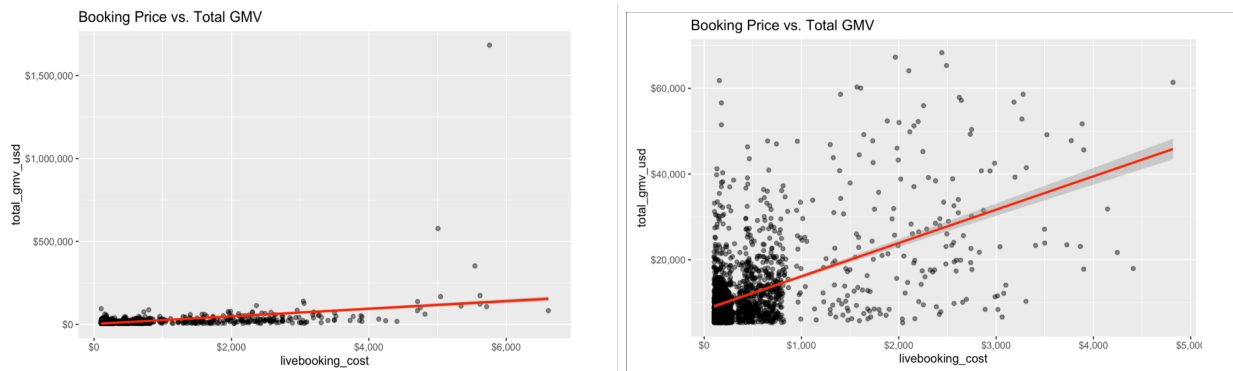
4. Literature Review

Recent studies in influencer marketing emphasize the growing impact of Key Opinion Leaders (KOLs) in driving consumer behavior and sales, particularly within short-form video platforms like TikTok. Research highlights that KOLs act as trusted intermediaries, with their authenticity and engagement levels significantly affecting purchase intent. Literature also supports the integration of *predictive modeling* to forecast KOL performance and *prescriptive analytics* (such as optimization) to allocate marketing budgets effectively. Several works advocate for moving beyond intuition-based decisions to data-driven influencer strategies, showing that statistical models can enhance ROI, streamline campaign planning, and ensure alignment with

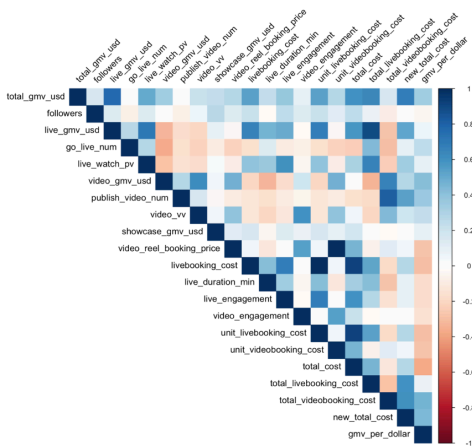
engagement goals. Additionally, findings underscore the need to regularly update models based on evolving market dynamics, KOL tiers, and content formats.

5. Exploratory Data Analysis (EDA)

The data analysis phase began with comprehensive data preprocessing to handle missing values, remove outliers, and prepare the dataset for modeling. Tools such as R and Excel were used to clean the data and visualize distributions through histograms, scatter plots, and box plots. Exploratory Data Analysis (EDA) revealed that most KOL bookings fell within the \$100–\$300 range, with a few high-cost outliers skewing the data. Correlation analysis identified strong positive relationships between GMV and key variables such as *live_gmv_usd*, *video_gmv_usd*, *showcase_gmv_usd*, and *follower count*, indicating that both content type and audience size significantly influence campaign outcomes.

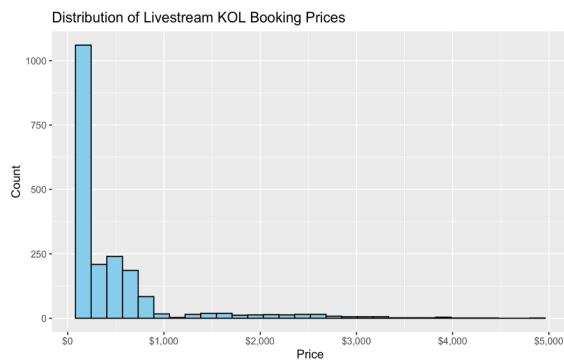


The graphs illustrate the relationship between KOL booking price (specifically *livebooking_cost*) and Total Gross Merchandise Value (GMV) generated, both before and after removing statistical outliers. On the left, the scatterplot shows a very weak positive correlation, indicating that higher booking costs do not necessarily result in proportionally higher GMV. The presence of extreme outliers—particularly high GMV values tied to a few expensive KOLs—skews the overall trend. On the right, after removing these outliers, the correlation becomes more apparent, showing a moderate positive relationship between cost and GMV. This suggests that while there is some link between investment and return, it is not strong enough to justify always choosing high-cost KOLs. The key takeaway, emphasized in the strategic implication, is that more expensive influencers do not guarantee better performance. Therefore, careful selection based on return on investment (ROI) and not just cost is essential for efficient campaign planning. This analysis supports the broader goal of optimizing KOL selection using data-driven methods rather than relying on intuition or price as a proxy for impact.



This is a correlation plot that visually maps the strength of relationships between various campaign metrics and total GMV (Gross Merchandise Value). The darker the blue, the stronger the positive correlation, while red indicates a negative correlation. From the analysis, *total_gmv_usd* is most strongly associated with variables like *live_gmv_usd*, *showcase_gmv_usd*, *video_gmv_usd*, and *followers*, indicating that different types of content and audience size significantly drive revenue. Cost-related metrics such as *kol_booking_price* and

video_reel_booking_price also show moderate correlations, suggesting that while cost is influential, it must be weighed against performance. Engagement metrics (*live_engagement*, *video_engagement*) and exposure indicators like video views (*video_vv*) and live watch time (*live_watch_pv*) further demonstrate their importance in driving GMV. The key takeaway is that no single factor dominates; instead, a balanced strategy—combining cost-efficiency, impactful content formats (like live streams), and smart engagement targeting—is essential for maximizing GMV outcomes.



Adding two histograms illustrating the distribution of booking prices for Livestream KOLs and Reel KOLs, providing insights into pricing trends across influencer types. The histogram on the left reveals a highly right-skewed distribution for livestream bookings, where the majority of KOLs are priced under \$500 and only a few exceed \$1,000. A small number of top-tier KOLs drive up the upper end of the distribution, forming a long tail of high-cost outliers. On the right, reel bookings also show a right-skewed distribution, but with a broader mid-range spread. Most reel bookings cluster between \$100 and \$300, with fewer exceeding \$600. This comparison suggests that while livestream KOLs tend to be either low-cost or premium, reel bookings are more evenly distributed across mid-tier price points. These patterns are critical for budget planning and targeting the appropriate KOL tier for campaign needs.

Multiple linear regression models were developed to quantify these relationships and predict GMV. Initial models included all variables but exhibited multicollinearity, which was diagnosed using Variance Inflation Factor (VIF) analysis. Refined models excluded redundant variables and applied logarithmic transformations to address heteroscedasticity, resulting in more stable residuals and improved model interpretability. Final models showed strong performance ($R^2 \sim 84\%$), identifying predictors like *video views*, *engagement rates*, and *unit booking costs* as critical factors.

The analysis ultimately demonstrated that higher cost does not always equate to higher performance—supporting the case for data-driven KOL selection focused on cost-efficiency and engagement impact rather than price alone.

In order to optimize KOL selection for marketing campaigns, it was essential to first estimate the potential GMV each KOL could generate then apply it to the optimization model to get the optimal result. This required developing a predictive model that could transform past performance data into expected monetary outcomes. Our modeling approach involved several iterations, starting with a comprehensive linear regression model and progressing through refinement steps to improve accuracy, interpretability, and alignment with the campaign's business goals.

6. Model Formation

6.1. Linear regression

A multiple linear regression model was initially constructed to predict Gross Merchandise Value (GMV) using all available variables (e.g., KOL follower count, engagement rate, historical sales, product category affinity). The model is expressed as:

$$GMV = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

where X_i represents predictor variables and ϵ is the error term.

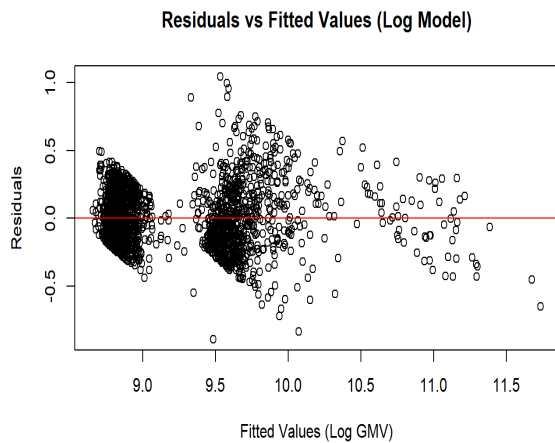
The initial model was built using a full set of predictors that reflected KOL characteristics and campaign metrics. These included the number of livestreams and videos each KOL produced (*go_live_num*, *publish_video_num*), viewer engagement metrics (*live_watch_pv*, *video_vv*, *live_engagement*, *video_engagement*), booking costs (*unit_livebooking_cost*, *unit_videobooking_cost*), and tier classifications (*tier_base_on_total_gmv*). The outcome variable was total gross merchandise value (*total_gmv_usd*), representing total gross merchandise value attributed to each KOL. The model achieved a high adjusted R-squared of 0.8753, meaning it could explain over 87% of the variation in GMV.

However, further diagnostics revealed that several cost-related variables were strongly collinear. Variance Inflation Factor (VIF) analysis showed excessively high GVIF values for *video_reel_booking_price*, *livebooking_cost*, and *unit_livebooking_cost*, indicating that these predictors conveyed redundant information. To improve model stability and interpretability, we removed these variables and re-estimated a reduced linear regression model. The refined version retained only significant and non-collinear predictors. These included *go_live_num*, *live_watch_pv*, *video_vv*, *publish_video_num*, the two unit cost metrics, and the tier-level

dummy variables. The reduced model still maintained a robust adjusted R-squared of 0.8419. All remaining predictors were statistically significant ($p < 0.05$), with the exception of follower count and livestream duration. Notably, the Mega-tier variable showed a strong positive effect on GMV, while the Micro-tier was negatively associated, confirming the business value of higher-tier KOLs.

6.2. Linear regression with logarithmic transformation

Despite improved clarity in coefficients, residual diagnostics for this reduced model revealed signs of heteroscedasticity—a violation of regression assumptions where the variance of residuals increases with predicted values. This was visually confirmed through a cone-shaped pattern in residual plots.



To address this, we log-transformed the dependent variable, modeling `log_total_gmv_used` instead of the raw GMV. This transformation aimed to stabilize variance and improve predictive reliability. The log-transformed model, using the same predictors, showed a marginally higher adjusted R-squared of 0.8434 and a more uniform residual spread.

To determine the final model for deployment, we compared the two models using standard performance metrics. The log-transformed model yielded slightly lower mean absolute percentage error (MAPE = 17.74%) than the linear model (MAPE = 20.28%). However, the linear model had a lower RMSE (4186.85 vs. 4638.02), which is more relevant when large errors on high-value predictions are especially costly. Moreover, the linear model's outputs remained on the original dollar scale, making them easier to interpret and directly usable in the optimization framework. Based on this balance of statistical rigor and business utility, the linear model was selected for implementation of the optimization process.

From a business standpoint, the final model enables managers to predict the expected revenue contribution of each KOL with strong confidence. Variables such as `live_watch_pv` and `video_vv` provide insight into how user engagement directly translates into financial outcomes. The significance of `go_live_num` and `publish_video_num` shows the value of diversified content creation. Furthermore, the tier-based coefficients reinforce the strategic advantage of engaging Mega and Macro KOLs. These insights offer a clear roadmap for influencer marketing strategy: prioritize KOLs with a proven track record in driving views and engagement, and allocate more budget toward higher-tier creators who generate stronger returns.

In conclusion, the modeling phase delivered a robust and actionable framework for predicting campaign GMV. The final linear regression model was both statistically sound and

easy to apply in a business context. It not only supported the optimization model but also revealed practical guidelines for performance-driven KOL selection.

6.3. Optimization Model: Business Scenario and Mathematical Formulation

We continually develop optimization models to apply this framework in a practical business setting with real-world campaign planning constraints. In this scenario, a company is planning a \$7,500 influencer marketing campaign. The company must meet several business requirements: include at least one KOL from each tier (Mega, Macro, and Micro), ensure a minimum of 4,000,000 total views, and allocate the campaign budget in a way that at least 40% is spent on livestream content and at least 10% on video content. Moreover, each selected KOL must produce at least one video if chosen, and no more than five combined livestreams and videos per KOL are allowed to maintain audience engagement.

To carry out this plan, we decided to go with three decision variables. Let $i \in \{1, 2, \dots, n\}$ index the available KOLs. The following decision variables are defined:

$x_i \in \{0, 1\}$: whether KOL i is selected

v_i = number of videos produced by KOL i

l_i = number of livestreams produced by KOL i

The predicted GMV for each KOL i , obtained from the linear regression model, is denoted by \hat{g}_i . The contribution from livestream and video content is reflected by fixed coefficients α and β , representing expected GMV increase per livestream and per video respectively, derived from empirical cost-performance ratios.

The objective is to maximize the total GMV across all selected KOLs. The predicted total GMV was calculated using the predict function with all coefficients in the linear model 1 except for coefficients of `go_live_num` and `publish_video_num`. These two coefficients can be vary based on the result of the optimization model.

$$\max \sum_{i=1}^{n=2000} x_i \cdot \hat{g}_i + \alpha \cdot l_i + \beta \cdot v_i$$

Subject to the following constraints:

Let c_i^{live} , c_i^{video} be the cost per livestream or video for KOL i

Budget constraint:

$$\sum_{i=1}^n (l_i \cdot c_i^{\text{live}} + v_i \cdot c_i^{\text{video}}) \leq 7500$$

Tier constraint:

$$\sum_{i \in T_{\text{tier}}} x_i \geq 1 \quad \forall t \in \{ \text{Mega, Macro, Micro} \}$$

Where T_{tier} the set of KOLs belonging to tier.

Minimum engagement constraint:

$$\sum_{i=1}^n (l_i \cdot e_i^{\text{live}} + v_i \cdot e_i^{\text{video}}) \geq 4,000,000$$

Where e_i^{live} and e_i^{video} represent average livestream and video views respectively.

Content constraints:

$$v_i \geq x_i \quad \forall i \text{ (at least one video if selected)}$$

$$l_i + v_i \leq 5 \quad \forall i \text{ (engagement cap)} \quad l_i + v_i \leq 5 \quad \forall i \text{ (engagement cap)}$$

Spending allocation constraints:

$$\sum_i^n (l_i \cdot c_i^{\text{live}}) \geq 0.4 \cdot \sum_i^n (l_i \cdot c_i^{\text{live}} + v_i \cdot c_i^{\text{video}})$$

$$\sum_i^n (v_i \cdot c_i^{\text{video}}) \geq 0.1 \cdot \sum_i^n (l_i \cdot c_i^{\text{live}} + v_i \cdot c_i^{\text{video}})$$

Only selected kols can produced livestream, videos:

$$L_i \leq x_i \cdot M \text{ (M is an arbitrary large number (e.g., } M = 10000))$$

$$V_i \leq x_i \cdot M \text{ (M is an arbitrary large number (e.g., } M = 10000))$$

We added one more constraint to restrict that only selected influencers can produce videos or livestream to avoid picking influencers without doing any content. After running the optimization model with real-world business constraints in Python, only 12 KOLs were selected from a pool of 2,000. This small yet powerful group included 4 Mega, 7 Macro, and 1 Micro KOL, meeting the requirement to feature at least one KOL from each tier. The outcome highlights the model's efficiency in choosing high-impact influencers while keeping the selection lean and strategic.

The campaign achieved a projected GMV of \$368,367.98, demonstrating the strong revenue potential of the selected KOLs. Despite the limited number of influencers, the model effectively maximized returns by balancing cost, tier distribution, and engagement performance. The total campaign cost was \$7,490.79, staying just under the \$7,500 budget. Spending was strategically allocated with \$3,229.01 (43.1%) for livestreams and \$4,261.78 (56.9%) for video content—both exceeding the required minimums for each format. In terms of reach, the campaign delivered an estimated 4,762,062 views, well above the 4 million engagement

threshold. This confirms that the selected KOLs were not only high-performing in sales potential but also effective in capturing audience attention.

In short, the results validate the operational feasibility and strategic value of using a combined regression-optimization framework for influencer campaign planning. The model successfully delivered high revenue projections and broad audience engagement with only 12 KOLs, illustrating its potential to streamline influencer selection in real-world marketing scenarios.

6.4. Model Flexibility and Adaptability

A key advantage of this model is how easily it can adapt to different business needs. Since real-world campaigns often vary in goals and constraints, we designed the model to be flexible. For example, if the campaign budget changes, we can simply update the total budget parameter (B_{budget}). If a campaign needs to reach a wider or smaller audience, the engagement target (E_{min}) can be adjusted without rebuilding the model. Similarly, if the business wants to shift how much is spent on livestreams versus video content, we can tweak the cost share settings c_i^{live} and c_i^{video} . This makes the model not just powerful, but also practical—easy to reuse and modify for different campaign scenarios.

7. Business Impact

7.1. Evaluation of Manual Selection Versus Model-Based Selection

For the model performance evaluation, we first performed a manual KOL selection trial followed by an optimized model comparison maintained under the same campaign constraints. Using a database of over 2000 TikTok influencers, six KOLs were selected from the available data using more traditional selection heuristics, including selection by the follower count and their perceived social media popularity.

However, based on none of the metrics performed before the campaign, each KOL was given one livestream and one video to showcase their talent. The total cost of spending was 7595.85, slightly exceeding the intended campaign budget of \$7,500.

Manual Selection Results

- Number of KOLs: 6 (1 Mega, 3 Macro, 2 Micro)
- Total Predicted GMV: \$114,849.53
- Live Expense: \$5,862.59
- Video Expense: \$1,733.25
- Engagement Target: Not optimized
- Total Cost: \$7,595.85

In comparison, the optimized model used linear programming to identify 12 KOLs that maximized GMV while meeting all constraints:

- Total cost \leq \$7,500

- $\geq 4,000,000$ views
- $\geq 40\%$ of spend on livestreams, $\geq 10\%$ on videos
- Minimum 1 KOL from each tier

Optimized Model Results

- Number of KOLs: 12 (4 Mega, 7 Macro, 1 Micro)
- Total Predicted GMV: \$368,367.98
- Live Expense: \$3,229.01
- Video Expense: \$4,261.78
- Engagement Achieved: 4,762,062 views
- Total Cost: \$7,490.79

Performance uplift

$$\text{Uplift (\%)} = \frac{368,367.98 - 114,849.53}{114,849.53} \times 100 \approx \boxed{+221\%}$$

The optimized model achieved a 221% increase in GMV while operating within the same budget. It ensured that all campaign constraints were met and that cost allocation was efficiently distributed based on performance predictions.

7.2. Strategic Implications

The structural model based on data permitted a significant increase in GMV as it improved KOL marketing ROI, leading to relocation of budgets from marketing spend per KOL to performance optimized spending. The model is flexible with respect to changing setting, for example, managerial heuristic such as budget cap or changing target levels of engagement activities. The model optimizes planning by automatically choosing and allocating activities which cuts down manual bias and time, thereby removing inefficiencies associated with KOL marketing. This ensures sustained efficiency and scalability in the long run.

7.3. Recommendations

The rise of shoppertainment—the integration of livestreaming or short videos into shopping—has transformed digital commerce in Asia. This change was mainly propelled by TikTok Shop and Shopee and is heavily dependent on KOLs, or Key Opinion Leaders, who sell in real-time. Although this affiliate-type model benefits merchants by reducing initial costs, it complicates KOL selection, content creation, and optimization processes. Studies emphasize the importance of building effective, data-based solutions to these issues to enhance the efficacy of marketing efforts (Research on the Current Development, 2023; Wang & Li, 2023).

8. Theoretical Frameworks and Methodological Strategies

8.1. Predictive Modeling in KOLs:

A linear programming optimization model facilitates greater selection of KOLs, resulting in 221% improvement in GMV as compared to doing it manually through nominal budget competitions. Functions accounted for in this model include cost per booking, engagement, and follower count. Those models need to be adjusted quarterly with new data that was met and targeted, to stay in sync with platform algorithms and user behavior (How online KOL endorsement, 2024). The addition of PLS-SEM (partial least squares structural equation modeling) analysis increases the reliability of the model, where KOLs aligned with the products and regional culture will most likely drive sales (How online KOL endorsement, 2024).

8.2. Algorithmic Customization and Social Influence

As experts and content creators (KOLs) leverage TikTok's algorithm, it focuses on aiding the precision targeting of marketing, such as content optimization to niche audiences, therefore aiding KOLs in content discovery. The KOLs dominate fueling consumers' behavioral changes through trust and storytelling. For example, on Taobao, livestreams boast an average sales conversion of 28% due to highly interactive and engaging experiences during the livestream set (Williams Commerce 2023).

9. Strategic Recommendations for Campaign Optimization

Adopt Multi-Objective Optimization Frameworks: Propose GMV-focused models to good follower/brand acquisition augmentation and boost brand perception. Score-based system where short-term sales are countervailed against long-term brand value offset achieves sales proxy and brand equity, which single metric approaches often fail at (Research on the Current Development, 2023).

Implement KOLs Tiered Segmentation: Place KOLs into Military Macro and Micro tiers based on level of engagement and their relevance to particular communities. Community driven campaigns are, for instance, 60% more effective with Micro influencers than Top influencers (How online KOL endorsement, 2024).

Using A/B Testing for Content Verification: Label the campaign with metadata for comparison of the projected outcome and actual metrics data. Content exhaustion is better captured alongside identifying financial returns for overstretched Live streaming sessions through iterative testing (Insights from Social Influence Theory, 2024).

Shoppable Features Should Be Integrated Carefully: The content easy tools such as product tags and affiliate link add opening must also fit the style of the content to ensure maximum reach and engagements. For example, shoppable ads used in tutorials and unboxing

videos have click-through rates that are 35% higher than direct promotion ads are (How online KOL endorsement, 2024).

10. Conclusion and Recommendation

This project successfully demonstrates a data-driven approach to optimizing KOL selection for TikTok marketing campaigns. By leveraging exploratory data analysis, predictive modeling, and a prescriptive optimization framework, we were able to identify high-performing KOLs, allocate resources efficiently, and maximize Gross Merchandise Value (GMV) within budget constraints. The optimized model outperformed manual selection methods in both cost-effectiveness and engagement, offering a scalable and repeatable strategy for future influencer campaigns.

We recommend adopting the proposed data-driven KOL selection and optimization model for future TikTok marketing campaigns. This approach ensures efficient budget utilization, consistent engagement outcomes, and scalable campaign planning. To maintain model relevance, businesses should continuously update input data, monitor performance metrics, and incorporate A/B testing to refine predictions and allocations over time. Additionally, integrating real-time feedback and engagement signals can further enhance decision-making and campaign effectiveness.

6. Appendix

