The Vibe-Aware Pricing Engine: A Predictive Analytics Approach to Optimizing Airbnb Revenue

MIS5460 Fall 2025 Group 1
Nicholas George, Sahil Medepalli, and Heath Verhasselt
georgen@iastate.edu / sahilmed@iastate.edu / heathv@iastate.edu

Heilmeier Summary: What we will do and why it matters

Objective, what are we trying to do: Learn a transparent pricing aid for Airbnb hosts that increases expected monthly revenue by combining standard listing features with a new Neighborhood Vibe Score engineered from guest review text.

Today and limits: Hosts use manual pricing or black box tools. These choices often miss hyper local demand signals and make it hard to justify price decisions.

Our approach and why it will work: Build the Vibe Score from review text using TF IDF, compact it with SVD, cluster neighborhoods into vibe segments, then train non linear models to predict 30 day occupancy. Use the model inside a simple price versus revenue simulator so a host can see clear trade offs.

Who cares: Hosts, property managers, and analysts who want data driven price guidance that they can explain.

Impact: Better price setting and clearer recommendations can raise revenue and reduce trial and error.

Risks and payoffs: Risk, the Vibe Score may be noisy or biased by sparse reviews. Payoff, a reusable, multi city method that turns unstructured text into a useful business feature and a tool that non technical users understand.

Resources: Inside Airbnb for three cities, Python notebooks, typical laptop hardware. We will pre-aggregate text and sample by month to keep memory use reasonable.

Timeline: Six weeks, finish by November 17. Week 1 data, Week 2 features and Vibe Score, Week 3 models, Week 4 interpretation and visuals, Week 5 draft, Week 6 polish.

Measuring progress and success: Technical: reduce MAE and RMSE over a baseline without Vibe. Business: improve expected monthly revenue on a fixed price grid by at least 5% versus the baseline simulator, and deliver a short pricing recommendation for three example listings.

Contents

1 Project Overview

The short-term rental market, dominated by platforms like Airbnb, operates with a significant information asymmetry. Hosts, ranging from individual homeowners to professional property managers, often lack the sophisticated, localized tools required for dynamic price optimization. This project addresses this market inefficiency by proposing the development of a data-driven tool designed to empower Airbnb hosts with optimal pricing strategies to maximize revenue potential.

The core mission is to construct an intelligent pricing engine that moves beyond rudimentary heuristics and opaque platform suggestions. The principal innovation of this project is the integration of a novel "Neighborhood Vibe Score"—a quantitative metric engineered from the unstructured text of guest reviews. This score is designed to capture the intangible, yet crucial, character of a location that heavily influences guest choice and willingness to pay.

The approach leverages the rich, multi-city Inside Airbnb open dataset to train and validate a suite of machine learning models. By combining traditional, structured listing features (e.g., property size, amenities) with the engineered Vibe Score, the resulting system will deliver transparent, data-driven, and location-aware pricing recommendations. The project's value lies not merely in prediction, but in the quantification of a qualitative concept. By transforming the subjective "vibe" of a neighborhood into a numerical feature, its economic impact on rental revenue can be measured, providing hosts with a previously unavailable business lever.

2 The Airbnb Pricing Optimization Challenge

The current landscape of Airbnb pricing is characterized by suboptimal strategies that create significant financial risk and uncertainty for hosts. A robust business case for a more advanced solution is evident when examining the limitations of existing practices.

2.1 Current Practices & Limitations

Hosts predominantly rely on two methods for setting prices: manual trial-and-error and Airbnb's native "Smart Pricing" tool. Both approaches have profound limitations.

- **Manual Pricing:** This strategy is fraught with risk. Hosts who underprice their properties leave potential revenue unrealized, while those who overprice face low occupancy rates and inconsistent cash flow. This high-effort, high-risk approach requires constant market monitoring and often yields suboptimal financial returns.
- **Airbnb's "Smart Pricing":** While an improvement over manual guesswork, this tool functions as a "black box." Its recommendations are generated by an opaque algorithm that is generic by design, failing to account for the unique attributes of a specific property or the nuanced, hyper-local demand drivers that define a neighborhood. Furthermore, the tool is primarily calibrated to optimize for bookings, a goal that does not always align with a host's objective of revenue maximization.

2.2 The Market Opportunity

A clear gap exists in the market for a third-party tool that offers transparency, customization, and deeper, more holistic insights. The opportunity lies in leveraging publicly available data to construct a model that is more attuned to the qualitative factors influencing traveler decisions. Current tools can compare properties based on the number of bedrooms, but they cannot differentiate between a property in a quiet, family-friendly area and one in a vibrant district known for its nightlife—even if both are in the same zip code. This project aims to fill that gap. The single market of a city is, in reality, a collection of distinct micro-markets defined by neighborhood character. By creating a "Vibe Score," this project performs a data-driven market segmentation of the neighborhoods themselves, allowing the predictive model to learn the different price elasticities of demand for each "vibe segment."

2.3 The Central Hypothesis

This project is founded on a central, testable hypothesis: *A property's revenue potential is significantly influenced not only by its physical attributes and price but also by the quantifiable 'vibe' of its surrounding neighborhood, as derived from collective guest sentiment.*

3 A Data-Driven Solution: The 'Vibe-Aware' Pricing Engine

The proposed solution is a multi-stage analytical framework that transforms raw data into an actionable pricing simulation tool. The methodology is designed to be robust, scalable, and grounded in established data science principles.

3.1 Foundational Data and Feature Engineering

The project will be built upon the Inside Airbnb open dataset, a comprehensive repository of public information from the Airbnb site at https://insideairbnb.com/get-the-data/. Specifically, the `listings.csv`, `calendar.csv`, and `reviews.csv` files for chosen cities will be utilized.

To ensure the model is generalizable and not biased by the idiosyncrasies of a single market, the analysis will be conducted on data from three diverse cities: New York City, USA; Austin, USA; and London, UK. This selection represents a global hub, a technology and event-driven market, and a historic, tourist-centric capital, respectively.

The primary predictive target will be an engineered metric, `occupancy_rate_30d`, calculated from the `calendar.csv` file. This represents the percentage of the next 30 days a listing is booked, offering a more stable and forward-looking target variable than historical occupancy.

A robust feature set will be engineered from the `listings.csv` data, including:

- **Property Attributes:** `room_type`, `accommodates`, `bathrooms`, `bedrooms`.
- **Host-Related Features:** `host_is_superhost`, `host_listings_count`, `host_identity_verified`.
- **Review-Based Features:** `number_of_reviews`, `review_scores_rating`, `reviews_per_month`.
- **Engineered Features:** `amenities_count` (derived by parsing the `amenities` text field) and `days_since_first_review` (as a proxy for listing age and experience).

3.2 Quantifying the Intangible: The Neighborhood 'Vibe Score'

This component represents the core innovation of the project, converting the unstructured text within `reviews.csv` into a powerful quantitative feature. The methodology follows a systematic, multi-step process grounded in natural language processing and information retrieval techniques.[1]

1. **Text Preprocessing:** All guest reviews will be aggregated at the neighborhood level. Standard NLP cleaning techniques will be applied, including converting text to lowercase, removing common stop-words (e.g., "the," "a," "is"), and applying stemming or lemmatization to reduce words to their root forms.
2. **Vectorization using TF-IDF:** The Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme will be applied to the cleaned text corpus.[1] This technique creates a document-term matrix where each neighborhood is represented as a high-dimensional vector. The TF-IDF score, calculated as $w_{t,d} = (1 + \log\ tf_{t,d}) \times \log(N/df_t)$, assigns higher weights to terms that are frequent within a specific neighborhood's reviews but rare across the entire city's reviews.[1] This effectively surfaces uniquely characteristic words that define a neighborhood's "vibe," such as "quiet," "park," and "family" for one area, versus "bar," "music," and "nightlife" for another.
3. **Dimensionality Reduction with LSI:** The raw TF-IDF matrix is inherently high-dimensional and sparse. To create a more compact and semantically meaningful representation, Latent Semantic Indexing (LSI) will be applied via Singular Value Decomposition (SVD).[1] LSI helps address issues of synonymy by grouping related terms into underlying "concepts." This technique factors the original term-document matrix A into three matrices, $A_k = U_k \Sigma_k V_k^T$, effectively projecting the neighborhoods into a lower-dimensional "concept space".[1] This step moves the analysis from a simple "bag of words" to a more powerful "bag of concepts."
4. **Clustering and Score Assignment:** A clustering algorithm, such as K-Means, will be applied to the LSI-transformed neighborhood vectors. This will group neighborhoods into a predefined number of "vibe clusters" (e.g., $k = 5$). Each resulting cluster will be qualitatively analyzed by examining its most prominent terms and assigned an intuitive label (e.g., "Vibrant & Central," "Quiet & Residential," "Hip & Alternative"). Finally, each listing in the dataset will be assigned the categorical ID of its neighborhood's cluster, which serves as the final "Vibe Score."

3.3 Predictive Modeling for Occupancy and Revenue

With a complete feature set, including the novel Vibe Score, a selection of robust, non-linear regression models will be trained to predict the target variable, `occupancy_rate_30d`. Models such as Gradient Boosting Machines (e.g., XGBoost, LightGBM) and Random Forests will be prioritized, as they excel at capturing complex, non-linear interactions between features.

Model training will be performed using a rigorous cross-validation framework to ensure the results are stable and generalizable. Performance will be evaluated using standard regression metrics, primarily Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

A key part of this stage is to statistically test the project's central hypothesis. Feature importance techniques, such as SHAP (SHapley Additive exPlanations) values and permutation importance, will be used to precisely quantify the Vibe Score's contribution to the model's predictive accuracy.

3.4 The Integrated Pricing Simulator

The ultimate deliverable is not a static model but a conceptual framework for a dynamic "what-if" analysis tool. The trained predictive model will serve as the engine for this simulator.

The functionality will allow a host to input their property's key features. The tool will then iterate through a range of potential nightly prices, feeding each price point into the model to generate a corresponding predicted occupancy rate. The final output will be a visualization of the revenue curve, plotting `Predicted Revenue (Price * Occupancy * 30)` against `Nightly Price`. This allows the host to visually and intuitively identify the revenue-maximizing price point for their specific property and its unique neighborhood context.

4 Anticipated Impact and Strategic Value

The successful execution of this project will generate significant value for a diverse range of stakeholders, extending far beyond individual hosts.

4.1 For Airbnb Hosts

- **Direct Financial Gain:** The primary impact is the ability to move from intuitive guesswork to data-driven pricing, enabling hosts to systematically identify and set prices that maximize their monthly revenue.
- **Competitive Benchmarking:** The tool provides a more meaningful way to benchmark a property. Instead of comparing only against similarly sized listings, hosts can compare their performance against properties with a similar "vibe," leading to more effective competitive positioning.

4.2 For Real Estate Investors & Analysts

- **Enhanced Property Valuation:** The "Vibe Score" introduces a new, quantitative metric for assessing the short-term rental potential of a property or an entire neighborhood, adding a layer of sophistication to investment analysis and valuation models.
- **Market Trend Identification:** By analyzing how neighborhood vibes evolve over time, investors can identify emerging, high-potential areas before their appeal is fully reflected in traditional real estate price indices.

4.3 For Urban Planners and Tourism Boards

- **Economic Impact of Culture:** The project provides a framework for measuring the tangible economic value of a neighborhood's intangible character and atmosphere, offering a new lens for understanding local tourism economies.
- **Policy Simulation:** The model could be used to understand how urban development projects, zoning changes, or new business openings that alter a neighborhood's "vibe" might impact local tourism revenue and housing dynamics.

This project also creates a framework for a new class of "psychogeographic" economic indicators. The Vibe Score is a snapshot of current public perception, which is dynamic. By applying this analysis to sequential data releases from Inside Airbnb, it becomes possible to track the evolution of a neighborhood's perceived character. A negative shift in vibe, as

captured by the language in reviews, could serve as a leading indicator of a future decline in tourism revenue, offering predictive power that precedes official economic statistics.

5 Project Execution and Success Measurement

A structured and realistic project plan is essential for delivering on the project's objectives within the allotted academic timeframe. This plan includes a detailed timeline, specific performance indicators, and a proactive approach to risk management.

5.1 Project Timeline and Milestones

The project will be executed over a structured six-week period. The following table provides a clear, week-by-week roadmap, breaking the project into discrete, achievable phases with specific deliverables to ensure consistent and trackable progress.

Week	Key Activities & Milestones	Deliverable(s)
1	Data Acquisition (3 cities), Merging & Cleaning (`listings`, `calendar`, `reviews`); Project Environment Setup	A single, clean, and merged dataset ready for analysis.
2	Exploratory Data Analysis (EDA); Feature Engineering (structural features); Develop & Validate "Vibe Score" Methodology	EDA notebook with initial visualizations; Engineered feature set including the Vibe Score.
3	Predictive Model Training (XGBoost, RF); Hyperparameter Tuning using GridSearch/RandomSearch; Initial Model Evaluation	Trained and tuned baseline models; Cross-validation performance metrics (MAE, RMSE).
4	Model Interpretation (SHAP plots); Development of Visualizations (Vibe Map, Simulator Mock-up); Begin drafting report methodology.	Feature importance plots; Draft of visual deliverables; Methodology section draft.
5	Final Analysis & Interpretation of Results; Draft Final Report and Presentation Slides	Complete draft of the final report and presentation submitted for feedback.
6	Incorporate Feedback; Finalize and Polish Report, Presentation, and Code Notebooks	Final project submission package (Report, Slides, Code).

Table 1: Project Timeline

5.2 Key Performance Indicators (KPIs)

Success will be measured against a combination of quantitative and qualitative criteria to ensure the project is not only technically sound but also practically valuable.

5.2.1 Quantitative Success

- **Model Performance:** The final predictive model must achieve a Mean Absolute Error (MAE) on the `occupancy_rate_30d` target that is at least 15% lower than a baseline model trained on the same data but excluding the "Vibe Score" feature.
- **Vibe Score Significance:** The engineered "Vibe Score" must rank within the top 50% of all features by importance (as measured by SHAP values or permutation importance) in the final model.

5.2.2 Qualitative Success

- **Interpretability:** The generated neighborhood "vibe" clusters must be qualitatively distinct and make intuitive sense upon manual inspection of their top-ranking TF-IDF terms.
- **Actionability:** The output of the conceptual pricing simulator must be clear, intuitive, and provide a direct, actionable pricing recommendation for a non-technical host.

5.3 Potential Risks and Mitigation Strategies

Proactive identification and planning for potential challenges are critical to project success.

- **Risk 1: Inaccurate "Vibe Scoring."** The NLP-derived score may fail to accurately capture the true character of a neighborhood, leading to a meaningless feature.
  - **Mitigation:** The Vibe Score will undergo qualitative validation. The top characteristic terms for each cluster will be manually reviewed for coherence. Furthermore, the clusters will be cross-referenced with external data sources, such as neighborhood guides or Yelp business category distributions, to establish face validity.
- **Risk 2: Data Sparsity.** Certain neighborhoods may have an insufficient number of reviews to generate a statistically stable Vibe Score.
  - **Mitigation:** A minimum threshold for the number of reviews will be established to profile a neighborhood. For neighborhoods that fall below this threshold, an imputation strategy will be employed, such as assigning them the Vibe Score of the nearest geographical neighbor that meets the data requirement.
- **Risk 3: Confounding Variables.** An observed correlation between "vibe" and revenue could be spurious, driven by an unobserved variable such as average neighborhood wealth or proximity to public transit.
  - **Mitigation:** The `price` feature will be included in the model to help control for the general cost level and socioeconomic status of a neighborhood. While it is impossible to eliminate all potential confounding variables, acknowledging this limitation and controlling for the most obvious ones will strengthen the validity of the analysis.

6 Envisioned Visual Deliverables

A primary goal of this project is to produce unique and insightful visual products that effectively communicate the analytical findings to both technical and non-technical audiences. The following three key visual deliverables are proposed.

6.1 The Multi-City 'Vibe Map'

- **Description:** An interactive choropleth map, developed using a library such as Plotly or Folium, will display the three selected cities. Neighborhood boundaries will be colored according to their assigned "Vibe Score" cluster. Users will be able to hover over any neighborhood to reveal a tooltip containing its name, its assigned vibe label (e.g., "Vibrant & Central"), and the top five TF-IDF keywords that define that vibe.
- **Analytical Value:** This visualization enables powerful cross-city comparisons. It can reveal a global typology of urban neighborhoods that transcends geography—for instance, demonstrating that Brooklyn's Williamsburg, London's Shoreditch, and Austin's East Austin may all belong to the same "Hip & Alternative" vibe cluster, suggesting common drivers of demand in these otherwise disparate locations.

6.2 The Interactive Pricing & Revenue Simulator

- **Description:** A dashboard mock-up, prototyped using a framework like Streamlit or Dash, will serve as the user interface for the pricing engine. A user can select a neighborhood (which automatically inputs the Vibe Score) and other property details. An interactive price slider will allow them to adjust the nightly rate, and a dynamic chart will instantly update to display the predicted occupancy and the full monthly revenue curve, clearly highlighting the optimal, revenue-maximizing price point.
- **Analytical Value:** This deliverable transforms the predictive model from an academic exercise into a practical, tactical decision-making tool. It makes the complex model output immediately intuitive and actionable for a host, democratizing access to sophisticated analytics.

6.3 SHAP Feature Importance Dashboard

- **Description:** A set of visualizations based on SHAP (SHapley Additive exPlanations) values will provide deep transparency into the model's decision-making process. A global summary plot will rank all features, including `price`, `accommodates`, and the `Vibe Score`, by their overall impact on predictions. Local, instance-level waterfall plots will deconstruct an individual prediction, showing exactly how each feature contributed to the final occupancy estimate for a single listing.
- **Analytical Value:** This moves beyond a simple "feature importance" list to explain the *why* behind the model's predictions. It can reveal complex feature interactions, such as showing that being a `host_is_superhost` has a larger

positive impact on occupancy in a ”Quiet & Residential” neighborhood (where trust is paramount) than it does in a ”Vibrant & Central” area (where location may be the dominant factor). This level of interpretability is crucial for building user trust and generating deeper business insights.

Team Contributions

Member	Planned contributions
Nicholas George	Data ingestion and cleaning, review text pipeline and Vibe Score construction, monthly snapshot alignment, storage format decisions
Sahil Medepalli	Baseline models, hyperparameter tuning, control function or double ML step for price, SHAP analysis
Heath Verhasselt	Visuals and simulator prototype, documentation and business write up, risk checks and validation, final integration

Table 2: Planned team contributions