Predictive Pricing Model—Group 12 Project Report

Authors: Shikhar Shah, Sudha Swain, Pranav Venkatadhri Pandalkudi Balaji, Jonathan Hansen, and Tiwari Durgesh
Affiliations: Indiana University - Luddy School of Informatics, Computing, and Engineering, Opinion Route, LLC

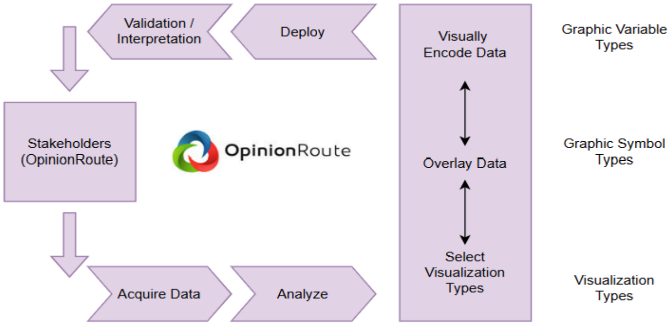

Fig. 1. The visual abstract figure shows the data processing of the project.

1. INTRODUCTION AND PRIOR WORK

This project's goal is to develop a pricing system to optimize the win rates of the client OpinionRoute. This will be done by leveraging the three datasets provided by the client. Our goal is to review these datasets and develop a comprehensive system of pricing that will allow the client to win more bid proposals. The team will aim to find the sweet spot of bid wins versus revenue (greater profits).

To determine pricing strategies, the team will implement statistical analysis, machine learning, and deep learning techniques to pull out more insights from the data. These different methods will aim to provide the client with a better understanding of the underlying connections between variables in the datasets and how they affect the overall pricing.

Prior work:
Example 1: A study that simulated the framework for construction bidding using an algorithm to determine optimal bidding decisions over time. The principle of developing a system to optimize bid wins would apply to our use case (link here).

Example 2: This research addresses the challenge of adjusting insurance renewal prices. It is framed as a sequential decision problem that uses a Markov Model Framework called Markov Decision Process and reinforcement learning. This study mirrors our challenge of finding a “sweet spot” between bid wins and profitability (link here).

The client has provided three datasets that have multiple features related to project bids and invoiced projects.

1.1 Stakeholder Groups

Sales and Pricing Analysts may want insights into the pricing factors and win rate trends to further develop the current pricing model.

Management and strategy planners would have a desire for interactive dashboards and more high-level analysis to help make ongoing decisions on market trends and company profitability.

Market researchers may want to understand how the pricing model affects the client and project outcomes.

1.2 Stakeholder Needs

The sales teams and pricing analysts will need data to support the pricing strategies and increase the win rates of bid proposals.

The management team will require ongoing analysis of pricing strategies that clearly show bid-win-loss ratios. Additional information for the management and sales teams would include depicting geographic distributions and relationships of key variables.

Strategy planners and management may additionally be interested in how a defined pricing strategy leads to downstream outcomes. For example, how does the existing pricing strategy affect long-term relationships or company profitability?

2. DATA ACQUISITION

The project will analyze three main datasets. The datasets were extracted from two different ERP systems at OpinionRoute, covering January 2024 through September 11th, 2024. The datasets include invoiced (successful bids) and unsuccessful bids, along with client segmentation data. The data was provided through a shared Google Drive folder, and assistance from the client has helped in understanding the data relevance, usage, and attributes.

Both the invoiced jobs and lost bids datasets span the same general time frame, although specific bid dates were not consistently available. Nonetheless, the datasets were deemed sufficiently accurate, relevant, and complete for the objectives of this project.

2.1 Data Sources

There are three datasets provided by the client :

1. Invoiced jobs/Bids won (invoiced_jobs_this_year)
2. Proposals/Bids lost (Deal+Item+Report+LOST).
3. Segment Dataset (Account+List+with+Segment)

The direct link to the data is provided here.

2.2 Data Description, Quality, and Coverage

The ‘Bids won’ dataset initially contained 677 rows and 14 attributes. After cleaning and one-hot encoding, it has 677 rows and 87 attributes. The ‘Bids Lost’ dataset started with 3493 rows and 12 attributes, reducing to 2694 rows with 10 attributes after initial cleaning. The ‘Segment’ dataset contains 154 rows and 4 attributes.

Table 1: Data Dictionary

Field Name	Description	Data Type
Customer Rate (CPI)	Price charged per interview	Continuous
Incidence Rate (IR)	% of qualified respondents	Continuous
Client Segment Type	Client classification categories	Categorical
Respondent Type	Survey classification categories	Categorical
Length of Interview (LOI)	Average interview duration (min)	Continuous
Quantity (Qty)	# of completed or required surveys	Discrete
Label	Bid success indicator	Binary

2.3 Data Cleaning and Handling

Several key cleaning steps were undertaken to prepare the data for analysis:

- Missing Segment Data: The initial bid datasets lacked segment information, which was resolved by joining the external Segment dataset using account keys.
- Missing Respondent Type: The respondent type attribute was initially overlooked but later incorporated due to its significant influence on Customer Rate values.
- Null Values: Less than 5% of total records contained null values and were removed.
- Outliers: Approximately 6% of records were identified as outliers (based on IQR thresholds) and were removed.
- Zero Values: Around 10 records had zero values for critical fields such as Customer Rate, Quantity, or Incidence Rate. These were considered likely data entry errors or test entries and were excluded.

One noted limitation is that, because full date features were not consistently available, it was not possible to confirm that Won and Lost bids perfectly align in terms of timing. This introduces a minor risk regarding potential hidden seasonality or temporal trends affecting pricing or success rates. After these cleaning steps, the dataset used for analysis contained 2544 total entries with 536 Won bids and 2008 Lost bids.

2.3 Data Pipeline Flow

The data pipeline was built primarily in Google Drive using Python in Google Colab for the data cleaning and ML modelling. The output data was provided in a CSV file for visualization in Power BI. The process began with acquiring the three raw datasets. The initial cleaning involved handling nulls, outliers, and zero values. Key features like the label were engineered. The data was filtered based on client requirements (e.g., ‘consumer’ respondent type, segment type) before being used for visualization and analysis.

3. DATA ANALYSIS

The project will include three main types of data analysis:

1. Exploratory Data Analysis
2. Statistical Analysis
3. Machine Learning/Deep Learning Models

Exploratory data analysis/Statistical analysis aims to identify patterns, correlations, outliers, and other statistical factors in the data related to the proposal bid price.

Initially, a logistic regression model was trained to classify bids as a win or a loss based on features including Customer Rate, Quantity, IR, and LOI. Numerical features were also standardized, and the dataset was split into 70% training and 30% testing. The model showed an accuracy of around 81.5%. However, due to significant class imbalance, namely significantly more lost bids than won bids, accuracy was misleading. The AUROC score of 0.67 indicated the model struggled to effectively differentiate between winning and losing bids, primarily predicting the majority class (Lost bids).

Recognizing the limitation of the classification approach given the data imbalance and lack of B2B data in the won bids category, the strategy changed. The focus shifted from classifying bid outcomes to predicting the optimal Customer Rate for potential wins. Various regression models, such as Linear Regression, Polynomial Regression, a simple Neural Net, and other ensembling techniques, such as Stacking, were implemented using only the Won bids data to learn the relationship between project characteristics (features) and the successful customer rate. This trained model was then applied to the Lost bids dataset to predict what the customer rate should have been, according to the trained model. This predicted rate serves as a data-driven suggestion for optimizing pricing on similar future bids that might otherwise be lost.

Table 2: Model Summary Chart

Model	Accuracy (%)	Error
Random Forest Regressor	49.54	47.68(MSE)
Polynomial Regression (degree 2)	38.22	58.38 (MSE)

Linear Regression	40	10.15 (RMSE)
Ridge Regression	33.3	8.98 (RMSE)
XG Boost	18.9	9.90 (RMSE)
Neural Network (Nadam-2 Dense layers)	39.22	57.44 (MSE)

The random forest regressor achieved a moderate R-squared value of around 0.4954 and a "Mean Squared Value" of 47.6879. High MSE indicates that the actual customer rate might be far off from the predicted customer rate, and the R-squared value tells us that the model is able to grasp about 50% of the model variance. While these numbers might not sound great, even if the absolute numbers are not extremely accurate, the relative ranking can still be very useful for pricing strategies, acting as an exploratory or supportive tool for identifying the ideal customer rate. Given that the R-squared value is only about 50%, we suspect that some important factors are missing, such as competitor pricing, customer urgency, project complexity, etc., which we did not have access to.

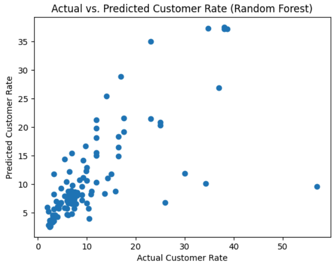

Figure 2. Actual vs. Predicted Customer Rates of the Bids won (Random Forest Regressor).

4. VISUALIZATIONS

The visualizations used were designed to provide insight into the relationship between project parameters, customer rate, and bid outcomes, with a primary focus on identifying potential optimal pricing to convert lost bids into won bids. The visualization approaches used allow stakeholders, particularly sales and pricing analysts, to explore how the project features include the predicted winning rate and then compare it to the historical actual rates.

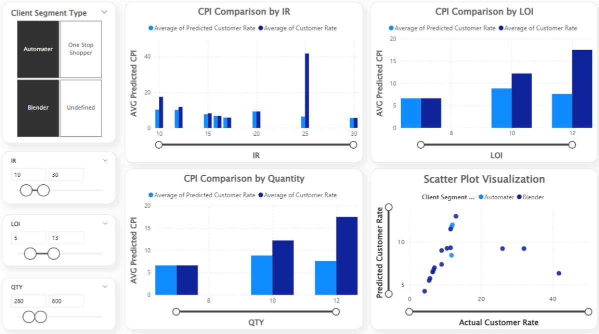

Fig. 3: Predicted vs Actual CPI and Parameter Comparisons link.

This interactive dashboard visualization allows stakeholders to explore the relationship between the predicted and actual customer rates based on key features within the dataset and the segment type.

The dashboard features include:

- Client Segment Filter: Allows data to be filtered by client segment (Automater, Blender, One Stop Shopper, and Undefined).
- Parameter Sliders for IR, LOI, and Qty: These sliders allow users to dynamically filter the data based on ranges of these parameters. This aligns with the client's request for bins for the ideal features.
- CPI Comparison by IR, LOI, and QTY: These bar charts show the average predicted CPI (from the random forest model) and the average actual customer rate from historical data, grouped by bins of IR, LOI, and QTY from the filters.
- Scatter Plot Visualization: This plot compares the predicted customer rate from the random forest regression model against the actual customer rate.

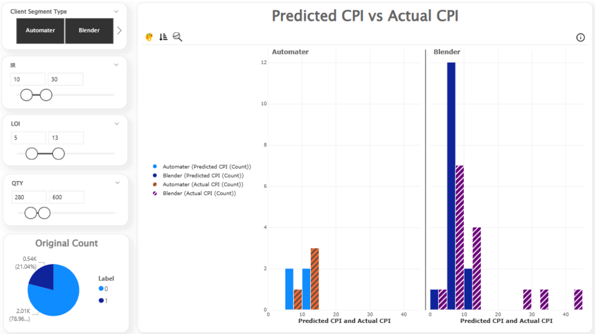
Client Segment Type
Automater
Blender
IR
LOI
QTY
Original Count
Label
Predicted CPI vs Actual CPI
Automater
Blender
Automater (Predicted CPI (Count))
Blender (Predicted CPI (Count))
Automater (Actual CPI (Count))
Blender (Actual CPI (Count))
Predicted CPI and Actual CPI

Figure 4: Predicted vs Actual CPI comparison between Segment type Automater and Blender link.

This visualization compares the distributions of the predicted customer rate from the random forest model applied to the lost bids and the actual customer rate from the historical data. The main point of this plot is to show the frequency distribution of each of these rates side-by-side. In this visualization, the same sliders and segment selections are available.

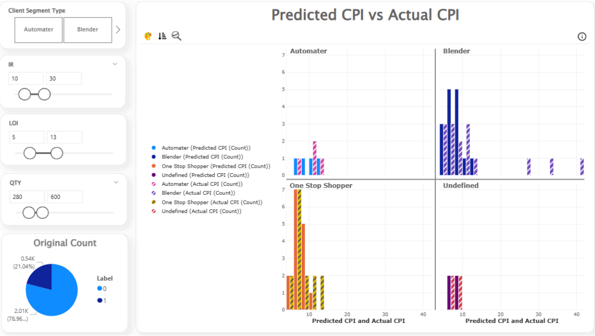

Figure 5: Predicted vs Actual CPI comparison between All Segment Type link.

5. USAGE AND CRITIQUE OF AI TOOLS

To have a better understanding of the dataset, its attributes related to market pricing, pricing factors that affect winning or losing the projects, and how they affect and relate to each other, we used AI tools for our knowledge gain. That's how we came up with these initial approaches.

- Gemini and ChatGPT were used for some of the data cleaning and ML coding tasks.
- Gemini Deep Research was used to find other similar work that has been used for predicting the best pricing.
- Gemini and ChatGPT were used to review work and suggest points to clarify within the text, or additional analysis that could be done with the data to enhance insights.

Limitations, such as biases, needed to be verified. Additionally, the models seemed to struggle with understanding the task to accomplish. These models served as assistants for coding, explanation, and structuring deliverables.

6. INTERPRETATION OF RESULTS

In our analysis, the visualization provides key insights into factors influencing OpinionRoute's bid outcomes and pricing strategies. Results vary based on variations of major features determining the pricing structure.

Figure 3 reveals how the predicted optimal customer rate from successful bids varies across different client segment types and ranges of IR, LOI, and Qty. The CPI comparison charts allow analysts to see that for some projects with a high Qty and high IR, the trained model suggests a lower Customer Rate compared to projects with low Qty and high IR. The scatter plot is specifically helpful in helping to identify the difference between historical pricing and predicted pricing.

For example, if a "Blender" segment bid with high LOI was lost at $30 Customer Rate (CPI), but the model predicted a winning rate closer to $20. This suggests that future bids with similar characteristics would be more likely to win at the predicted rate.

Figure 5 provides an aggregated view showing where the models predicted winning prices tend to fall relative to the historical. Overall, we see the predicted prices to be below that of historical prices, which reinforces the idea that OpinionRoute is overpricing the proposals per the signals from their own successful bids.

Combining the insights from these visualizations, stakeholders gain a better understanding of the sensitivity of pricing across project types and client segments. The predicted rates serve as data-informed benchmarks. While not a rule, proposing bid customers' rates closer to the predicted value from the random forest model will likely increase the bid acceptance rate.

Validation and Future Work:

While this approach does indicate a good starting point, further analysis should be done to include additional validation of models and data. Using a k-fold validation could further enhance the model and ensure the performance generalizes well across other sections of unseen data. While the dataset used was not small, it was not a large dataset either. Giving the machine learning model a larger dataset could help it learn more intricacy, although a dropout method may need to be used to ensure the model does not overfit.

Finally, another point of future work would be to ensure that profit is being maximized at any acceptance rate. Given additional data, the initial logistic regression model implemented could be a good starting point for determining the win rate for maximized profit.

ACKNOWLEDGEMENTS

We would like to show our gratitude to our professors Katy Borner and Michael Ginda for their guidance and training, which have been instrumental in shaping our visualization expertise. We are also thankful for the assistance from Teaching Assistant Divya Prasanth for his support and insights throughout the project. Lastly, we would like to thank OpinionRoute, Kayla Cousineau, and Terence McCarron for their support and insight into the data.

REFERENCES

1. OpinionRoute. (2024). *Invoiced jobs and bid proposals lost* [Dataset]. OpinionRoute. Retrieved March 6, 2025, from: https://drive.google.com/drive/folders/12zvOu9qpiO8AlQlUP2S-P2dd_q9PRCda

2. Assaad, R., Ahmed, M. O., El-Adaway, I. H., Elsayegh, A., & Nadendla, V. S. S. (2020). *Comparing the impact of learning in bidding decision-making processes using algorithmic game theory*. American Society of Civil Engineers. Retrieved March 8, 2025, from: https://doi.org/10.1061/(ASCE)ME.1943-5479.0000867

3. Krasheninnikova, E., García, J., Maestre, R., & Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering Applications of Artificial Intelligence*. Retrieved March 8, 2025, from https://e-archivo.uc3m.es/rest/api/core/bitstreams/03e94ce5-863b-4b3c-9deb-c9fa00eb1fe0/content

4. GitHub Repo containing Python files and PowerBI: https://github.com/sudswain-hue/Predictive-Pricing-Model---Grp-12