Capstone Proposal

Project Name: DKomplex: Intelligent Pricing Forecast Model

Course Name: CST 489 – 499: Computer Science Capstone

Client Name: Justin Loza

Student Names: Chika Starks, Daniel Arias, Deborah Shaw, Joshua Okero, Nicole Al-Sabah

Advisor Name: Dylan Kaiser

Term: Summer 2025

Executive Summary

The capstone project continues the work a previous student group started last semester with DKomplex, an IT firm based in Seaside, California. The company's goal is to build a machine learning model that predicts the Price Per Unit (PPU) most likely to lead to a successful revenue outcome for a specific plastic product. This is part of an endeavor to support DKomplex's pricing decisions by uncovering patterns in historical sales data, customer behavior, product quantities, and other characteristics that influence the pricing.

Currently, DKomplex has challenges in consistently calculating ideal pricing. Their data is pulled from multiple systems such as quotes, opportunities, and orders which have inconsistent data schemas. Other challenges include gaps in the data, inconsistent naming, and internal sales that make it difficult to analyze pricing patterns. In addition, DKomplex does not have abundant competitor pricing info to work with, which creates another layer of complexity. The model is being designed to take the historical sales data DKomplex has available and find patterns that show what types of prices are most likely to lead to a successful sale.

For the Summer 2025 capstone, the focus includes integrating timing features, such as economic factors and competitor data, to improve the model's applicability in realistic scenarios. The project also involves strict data privacy, NDAs and audits are required to ensure no data is exported outside of DKomplex's systems. The outcome of this work is to support DKomplex's internal leaders with data-driven pricing strategies that align with both technical needs and user adoption goals. The pricing strategies would support DKomplex's work with clients, including private companies, nonprofits, and government agencies. Our goal is to deliver meaningful progress that can be used by DKomplex and future student teams continuing this project.

Table of Contents

Introduction/Background

Project Name and Description

The project is titled "DKomplex: Intelligent Pricing Forecast Model" and is a continuation of a client based capstone project in partnership with DKomplex, an IT firm based in Seaside, California. The goal is to develop a machine learning model that predicts the Price Per Unit (PPU) most likely to lead to a successful revenue outcome for a specific plastic product. This model is intended to support pricing decisions by identifying patterns in historical sales data, including customer behavior, product quantity, and other relevant information. The project is important because pricing directly impacts profitability, and DKomplex is aiming to build a data-driven tool that can be improved to meet their business needs.

Problem and/or Issue in Technology

DKomplex is currently facing challenges due to inconsistent data schemas across various systems, such as quotes, opportunities, and orders. This inconsistency makes it difficult for the company to identify underlying patterns from historical data. In addition, the quality of the data is compromised by missing values and inconsistent naming conventions, which delays the development of a reliable price prediction model for forecasting future Price Per Unit (PPU) of a specific plastic product. Another contributing factor to pricing difficulties is the influence of internal sales, which can compromise the historical integrity of pricing data. Additionally, DKomplex lacks access to usable competitor data, making it even more challenging to determine appropriate market pricing. Together, these obstacles prevent the company from setting optimal prices that could maximize profitability.

Solution to the Problem and/or Issue in Technology

The solution begins with cleaning the historical data by enforcing consistent naming conventions and handling missing values to ensure alignment across datasets. This includes validating data by removing outliers and converting records into consistent schemas for quotes, opportunities, and orders prior to processing. Normalization will also be applied to improve the accuracy of price predictions. We will use common data science tools such as PySpark, pandas, and NumPy to handle and process the cleaned data. A regression model will be used, and its performance will be evaluated using statistical metrics such as the R2 score. Visualization tools like Matplotlib and Seaborn will help interpret the performance of the machine learning model. Also, since the company lacks usable competitor data, we may need synthetically generated training data, potentially through data augmentation, to enhance the model's predicting ability.

Environmental Scan/Literature Review

Machine learning models have been widely used for price predictions, with various algorithms available to help companies maximize their profits. Support Vector Machine (SVM), a supervised learning algorithm, is one of the commonly used models for predicting product prices. SVM is also suitable for regression tasks through its variant, Support Vector Regression (SVR), which allows users to define a maximum error or EEE (epsilon) within which predictors are considered acceptable. This flexibility enables businesses to estimate competitive market prices while maintaining a balance between accuracy and generalization. For example, SVM has been successfully used in Thailand to predict the prices of polypropylene granules, making it one of the good choices for our project (Liu & Zhang, 2023).

Tree-based machine learning algorithms are also strong candidates for price prediction tasks. These models work by splitting the dataset into subsets based on feature values, forming a tree-like structure to predict continuous outcomes. Tree-based methods can effectively estimate the cost of plastic products, such as injection-molded parts, by analyzing features like material properties, part geometry, and production parameters. In particular, decision tree regression and random forest regression have proven useful for uncovering relationships in pricing data. When combined with techniques like gradient boosting, these models can significantly improve accuracy and help businesses determine optimized price points based on historical and contextual data. Their interpretability and performance make them especially valuable in pricing strategies (Bernsteiner & Wimmer, 2023).

Stakeholders

This project has multiple stakeholders, starting with DKomplex, the primary stakeholder and project sponsor. DKomplex stands to gain a predictive pricing machine learning model that can support the business in making more accurate and data-driven pricing decisions. If the model's outcomes are inaccurate or difficult to implement, the result may lead to poor pricing decisions, reduced competitiveness, and a loss of client trust in the tool.

Developers, including our team, are responsible for building, testing, and optimizing the machine learning model. The developers stand to gain valuable experience in applying machine learning techniques to real-world business problems. The success of the data implementation not only enhances the developers' skills but also contributes to the overall success of the organization. If the model fails to meet business needs or deliver accurate predictions, it could result in poor pricing decisions, decreased competitiveness, and a loss of client trust, ultimately impacting the organization’s reputation and financial performance.

End users and clients, including private companies, nonprofits, and government agencies, will use our machine learning model and stand to gain data-driven pricing strategies to enhance decision making within their organizations. If the model’s predictions are inaccurate or misleading, it could negatively impact end users’ and clients’ performance and potentially lead to revenue loss.

Ethical Considerations

This project involves highly confidential data, such as invoiced sales records, account histories, and internal pricing strategies. Improper handling of this information could lead to breaches of privacy, competitive risks, and loss of client trust.

1. Privacy and Security:

It is essential to protect the business’s confidential information, such as sales data and pricing strategies. If this data is exposed beyond the project team, it could result in serious privacy breaches and competitive risks.

Mitigation: All team members are required to sign NDAs and will follow strict data security protocols to ensure that all information remains confidential and is accessed only within secure environments.

2. Accessibility: Project insights and tools should be accessible to all relevant stakeholders, including non-technical users or those with accessibility needs.

Mitigation: We will ensure that all reports, visualizations, and outputs follow usability and accessibility best practices.

3. Bias and Fairness: The historical data may contain biased patterns that could lead to unfair pricing recommendations.

Mitigation: Regular checks will be conducted to monitor data trends and model outputs, helping to identify and address any bias.

4. Transparency: Our machine learning results may not always be easily understood, which could potentially mislead end users.

Mitigation: We will provide clear labels, explanations, and documentation to reduce confusion and ensure users correctly interpret the results.

Legal Considerations

1. Non-Disclosure Agreements (NDAs):

All team members are required to sign NDAs as part of the engagement with DKomplex and its client(s). These agreements are essential to ensure that all sensitive and proprietary data accessed during the project remains strictly confidential and is retained within dKomplex's secure systems.

2. Copyrights:

Any deliverables, including models, code, and documentation, are the intellectual property of DKomplex or its client(s). All team members must ensure proper licensing for any third-party content and avoid using copyrighted materials without permission.

3. Data Usage Permissions:

Access to internal datasets, including invoiced sales, opportunity, and quote tables, is limited to the scope of this capstone project. All team members are responsible for ensuring data security and preventing unauthorized access or export.

Project Goals and Objectives

Goal (Long-Term)	**Objectives (Concrete, Measurable)**
Develop a robust, data-driven model for optimal pricing of plastic products	- Clean and standardize DKomplex’s historical sales datasets - Integrate relevant economic and synthetic competitor features - Build machine learning regression models and - Visualize results for technical and non-technical audiences evaluate their effectiveness
Deliver actionable insights for DKomplex’s leadership	- Document findings and recommendations in a comprehensive final report - Present results and methodology to client and future student teams
Ensure ethical, secure, and sustainable development	- Present results and methodology to client and future student teams - Follow strict NDAs and data privacy requirements throughout project lifecycle - Build clear documentation for future expansion or handoff

Final Deliverables

- **Cleaned and integrated sales dataset**, including all code and data documentation required for use.
- **Trained machine learning model(s)** for Price Per Unit (PPU) prediction, along with model evaluation summaries.
- **Visual dashboard or report** presenting key findings, model results, and feature importance analyses.
- **Comprehensive technical documentation** covering the data pipeline, code, methodology, and guidance.
- **Final presentation** to DKomplex and capstone advisors, summarizing results and recommendations.
- **Client sign-off/approval** upon acceptance and handoff of all deliverables.

Approach/Methodology

- **Agile/Iterative Development:** The team will use an agile approach, with weekly meetings, milestone reviews, and regular feedback from the client and advisor.
- **Research & Requirement Gathering:** Begin with a review of related academic and industry work, including analysis of prior DKomplex projects.
- **Data Engineering & Preparation:** Collect and clean internal sales, quote, and opportunity data, ensuring consistency and completeness.
- **Feature Engineering:** Develop features based on economic factors and simulate competitor data as needed.
- **Model Creation:** Build and test multiple regression models, including Support Vector Regression and tree-based techniques, and compare performance.

- **Visualization:** Use visualization libraries to interpret and communicate model results.
- **Documentation:** Maintain and update thorough records of all processes, code, and findings for current and future project teams.
- **Stakeholder Collaboration:** Maintain open communication through regular check-ins and milestone reviews with DKomplex and advisors.

Timeline and Resources

Timeline/Milestones Table

Milestone	Start	End	Description
Project Planning/Research	Week 1	Week 2	Literature review, goal setting, data/environment access
Data Collection/Cleaning	Week 2	Week 3	Data extraction, schema alignment, cleaning, merging
Feature Engineering	Week 4	Week 5	Adding economic/synthetic features, exploratory analysis
Model Development	Week 5	Week 7	Training, tuning, and selection of regression models
Evaluation & Visualization	Week 7	Week 8	Model assessment, creating visualizations and dashboards
Documentation & Handoff Prep	Week 8	Week 9	Prepare documentation, user guides, and handoff materials

Client Presentation & Handoff	Week 9	Week 10	Final presentation, client feedback, makings for future team

Resources Needed

- **Data Systems:** Secure access to DKomplex sales and related databases
- **Computing Resources:** Approved environment for data processing and modeling
- **Software & Tools:** Python (pandas, NumPy, PySpark, scikit-learn), Jupyter Notebooks, Matplotlib, Seaborn
- **Collaboration Tools:** GitHub for version control, Google Drive or MS Teams for documentation and storage
- **Documentation:** Google Docs/Sheets and Markdown for project artifacts and record-keeping
- **Security Protocols:** NDA agreements and secure data handling as per DKomplex requirements

Platform

- **Primary Platform:** Python (Jupyter, PyCharm) within the DKomplex secure environment.
- **Reason:** Python is industry-standard for data science, machine learning, and rapid prototyping, and complies with the client's security policies. Jupyter Notebooks allow interactive exploration and are ideal for collaborative work. GitHub will be used for code versioning and sharing.

Risks and Dependencies

Risks

Risk(Events/Conditions)	Mitigation
Delays in receiving/accessing DKomplex data	Early and ongoing communication with client, contingency planning
Data quality issues (missing, incorrect, uneven)	Implement robust cleaning processes, clearly document assumptions and gaps
Limitations of synthetic competitor data	Clearly communicate as a limitation in results/recommendations
Technical issues (software, environment)	Regular backups, version control, use of tested libraries
Changes in project requirements/scope creep	Frequent check-ins, firm agreement on deliverables
Team availability (holidays, illnesses, etc.)	Implement backup assignees if needed, buffer schedule

Dependencies

- Cannot develop or test machine learning models until data is provided and cleaned.
- Later milestones (modeling, visualization, reporting) depend on successful completion of foundational (data) tasks.

Testing Plan

- **Functionality Testing:** Peer review and advisor assessment of all code and processes; use sample/test data where appropriate.
- **Model Performance Testing:** Use statistical methods (e.g., R^2, MAE, cross-validation) to evaluate the accuracy and robustness of predictive models.

- **Usability Testing:** Share dashboards and technical documentation with the client for feedback. Adjust visualizations and documentation for clarity.
- **Expert Review:** Seek specific client input on interpretability, practicality, and actionable value of the outputs.

Team Members and Roles

Name	**Role(s)/Responsibility**
Chika Starks	Data integration, client communication, testing
Daniel Arias	Model development/tuning, code base management
Deborah Shaw	Project management, reporting, presentation lead
Joshua Okero	Data cleaning/engineering, documentation
Nicole Al-Sabah	Feature engineering, visualizations, documentation

Division of Labor:

All team members participate in literature review, weekly meetings, and strategic planning. A team member will be given a backup role in case of an absence. Final deliverables are collaboratively reviewed and tested before handoff to DKomplex.

References

Bernsteiner, R., & Wimmer, T. (2023). A Machine Learning Approach for Automated Cost Estimation of Plastic Injection Molding Parts. *Cloud Computing and Data Science*, 4(2), 87–111. https://doi.org/10.37256/ccds.4220232277

DKomplex. (n.d.). *DKomplex: The IT+ Firm*. Retrieved May 19, 2025, from https://www.dkomplex.com/

Liu, Y., & Zhang, H. (2023). Machine Learning-Based Price Forecasting for Polypropylene Granules in Thailand. *CIIS '23: Proceedings of the 2023 6th International Conference on Computational Intelligence and Intelligent Systems*, 14–19. https://doi.org/10.1145/3638209.3638212