Capstone Project Report

Krutika Dhananjay Deshpande

Spring 2024

—

Sales Forecasting: Private Kaggle Challenge

—

Under the guidance of Dr. Prof. Surendra Sarnikar

INTRODUCTION

In today's highly competitive business landscape, companies are constantly seeking ways to optimize their operations and gain a competitive edge. One such company, a steel manufacturer, faced a challenge in accurately forecasting its sales to various customers across different industries. Accurate sales forecasting is crucial for efficient inventory management, resource allocation, and effective planning, directly impacting the company's profitability and growth.

THE PROBLEM STATEMENT

The sales forecasting problem presented itself as a complex challenge due to the involvement of multiple factors.

The company's sales were influenced not only by customer-specific attributes, such as financial ratios, revenue growth projections, and market share changes, but also by broader economic indicators like consumer sentiment, interest rates, and purchasing managers' indices.

With a diverse customer base spanning the Automotive, Metal Fabrication, and Infrastructure industries, the company needed a robust solution that could accurately predict quarterly sales for each of its 75 customers. Precise sales forecasts would enable the company to make informed decisions, streamline operations, and better meet customer demand.

Accurately predicting sales required the ability to analyze and combine these diverse data sources effectively.

The goal was to develop a robust predictive model that could leverage the available data, including company-specific information and economic indicators, to forecast quarterly sales for each customer accurately.

THE APPROACH

The project required a comprehensive data science approach, involving several key steps:

1. **Data Collection and Exploration:** Gathering and understanding the company's customer data, financial metrics, and relevant economic indicators.

   Here, we were already provided with the Datasets to use.
   - train.csv - the training set
   - test.csv - the test set
   - sample_submission.csv - a sample submission file in the correct format
   - EconomicIndicators.csv - supplemental economic information for the corresponding period in train/test data

2. **Data Visualization and Exploratory Data Analysis (EDA):** To gain insights into the data and identify potential patterns or relationships, various data visualization techniques were employed. This included creating histograms, scatter plots, box plots, and correlation matrices to understand the distributions of variables, identify outliers, and assess the relationships between features and the target variable (quarterly sales). EDA techniques like summary statistics, missing value analysis, and analyzing categorical variable distributions were also utilized to thoroughly explore the data.

3. **Data Preprocessing and Feature Engineering:** Cleaning and transforming the data, handling missing values, and creating new features that could potentially improve the predictive power of the model. Examples of engineered features included ratios, interactions, polynomial features, binned

features, domain-specific features like average economic indicators.

4. **Model Development and Evaluation**: With the preprocessed and feature-engineered data, various machine learning models were explored for the sales forecasting task. Ensemble methods like Random Forests and Gradient Boosting were investigated, as well as neural networks, due to their ability to capture complex non-linear relationships. The models were trained on the historical data, and their performance was evaluated using appropriate metrics, such as Mean Absolute Error (MAE), on a held-out validation set.

5. **Hyperparameter Tuning and Model Selection**: To optimize the performance of the models, techniques like grid search or random search were employed for hyperparameter tuning. This involved systematically trying different combinations of hyperparameters (e.g., number of trees, depth of trees, learning rates) and selecting the configuration that yielded the best performance on the validation set. Cross-validation was used to ensure the robustness of the models and avoid overfitting.

6. **Model Deployment and Monitoring:** Once the best-performing model was selected, it was deployed into the company's systems for real-time sales forecasting. This involved integrating the trained model into the existing software infrastructure or developing a new application to serve predictions. Continuous monitoring of the model's performance was essential to ensure its accuracy and relevance over time. Periodic retraining or updating of the model might be necessary to account for changes in market conditions, customer behavior, or economic factors.

7. **Iterative Refinement**: The data science process is iterative, and refinements can be made at various stages to improve the model's performance. This could involve revisiting the feature engineering, exploring advanced techniques or something else.

THE SOLUTION

Data Preprocessing:

The datasets were preprocessed to handle missing values and prepare the data for analysis. The missing values in the InventoryRatio column were imputed using the median value. Rows with missing Sales values in the training dataset were dropped. The test dataset was checked for duplicate ID values, and the first occurrence was kept in case of duplicates.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis was performed to gain insights into the data and identify potential patterns or relationships. Key findings include:

- Distribution of numeric features: Histograms were plotted to understand the distribution of numeric features such as QuickRatio, InventoryRatio, RevenueGrowth, and MarketshareChange.
- Relationship between categorical features and Sales: Box plots were created to visualize the relationship between categorical features like Quarter, Bond_rating, and Stock_rating with the Sales variable.
- Unique values and counts for categorical features: The unique values and their counts were analyzed for categorical features such as Company, Quarter, Bond rating, Stock rating, Region, and Industry.
- Correlation analysis: The correlation between numeric features and the Sales variable was examined to identify potential linear relationships.

Feature Engineering:

Several feature engineering techniques were applied to create new features and capture additional information:

1. Interaction features:
    - QuickRatio_InventoryRatio: Created by dividing QuickRatio by InventoryRatio to capture the interaction between these two ratios.
    - RevenueGrowth_MarketshareChange: Created by multiplying RevenueGrowth by MarketshareChange to capture the interaction between revenue growth and market share changes.
2. Polynomial features:
    - QuickRatio_Squared: Created by squaring the QuickRatio feature to capture non-linear relationships.
3. Binning:
    - QuickRatio_Binned: The QuickRatio feature was binned into three categories (Low, Medium, High) based on predefined thresholds. This technique helps to capture non-linear relationships and reduces the impact of outliers.
4. One-hot encoding:
    - Company: The Company feature was one-hot encoded, creating binary variables for each unique company (e.g., Company_CMP01, Company_CMP02, etc.).
    - Region: The Region feature was one-hot encoded, creating binary variables for each unique region (e.g., Region_East, Region_North, etc.).
    - Industry: The industry feature was one-hot encoded, creating binary variables for each unique industry (e.g., Industry_Automobile, Industry_Infrastructure, etc.).
    - QuickRatio_Binned: The binned QuickRatio feature was one-hot encoded, creating binary variables for each bin (e.g., QuickRatio_Binned_Low, QuickRatio_Binned_Medium, QuickRatio_Binned_High).

5. Average economic indicators:
    - Avg_Economic_Indicators: Created by taking the average of various economic indicators to capture the overall economic conditions.

These feature engineering techniques were applied to both the training and test datasets to ensure consistency and to capture relevant information for the sales forecasting model.

Model Development and Evaluation:

After trying several models, Random Forest Regressor model provided me the best results for sales forecasting due to its ability to handle complex relationships and its robustness to outliers. The model was trained using the preprocessed and feature-engineered data, with hyperparameters set to n_estimators=8000 and random_state=42.

The data was split into training and validation sets using a train-test split with a test size of 0.2. The model was trained on the training set and evaluated on the validation set using the Mean Absolute Error (MAE) metric. The validation MAE obtained was 302.61, indicating the average absolute difference between the predicted and actual sales values.

The sales forecasting model has the potential to provide valuable insights and support decision-making processes for the company. By accurately predicting sales, the company can optimize inventory management, resource allocation, and planning, leading to improved efficiency and profitability. The model's predictions can be used to identify potential opportunities or risks, enabling proactive measures to be taken.

ITERATIONS & THEIR RESULTS

No.	What was done/changed	MAE	Kaggle Score
1	Used Label encoder for Categorical Variables and KNN imputer for Numerical variables. Model: Random Forest Regressor	472.22	1284
2	Used Ordinal Encoding for Bond & Stock rating. One Hot Encoding for rest of the Categorical Variables. Missing values in numerical variables were imputed with respective mean values. Same model as before.	85.36	920
3	Tried Scaling using StandardScaler(), rest was the same Model: Gradient Boost	90.69	958
4	Mapped Month to Quarter Merged Economic indicators. Used only variables with positive correlation coefficient. Model: Random Forest Regressor	75.81	776.90
5	Model changed: XGBoost	251.81	707.03
6	Tried Time sorting technique (by month) Model: Random Forest Regressor	319.97	712.95
7	Tried Time sorting technique (by month) and split at 14 instead of defining test & validation set. Model used same as before.	75	704.01
8	Imputed InventoryRatio by Median to handle missing values. Tried Sequential Feature Selection Model: Random Forest Regressor	89.63	923
9	Time sorted by month, and tried One hot encoding for all categorical variables. Removed Sequential feature selection and used feature importance instead. Went back to defining train, test and validation data for the model. Same model as before	44.275	716.69
10	Went back to Ordinal encoding to stock and bond rating variables and rest had One Hot encoding. Removed feature importance. Rest same as above. Added new features: • QuickRatio_InventoryRation = QuickRatio/Inventory Ratio • RevenueGrowth_MarketshareChange = RevenueGrowth*MarketshareChange • QuickRatio_Squared= QuickRatio ** 2 • Avg_Economic_Indicators = [economic_cols].mean(axis=1) Model: Random Forest Regressor N_estimators = 1000	294.45	667.62

USE OF GENERATIVE AI

During the course of this project, generative AI, specifically the AI language model developed by Anthropic, was utilized to assist in various aspects of the project. The AI model, named Claude, provided valuable insights, suggestions, and guidance throughout the project lifecycle. Some key areas where generative AI was employed include:

1. Code:
   By providing the idea that I wanted to implement, I gave the details to Claude, and asked for the Python code to implement it.
   Of course, the code needed manual adaptation to match the variable names and the workflow to fit my solution.

2. Exploratory Data Analysis (EDA):
   Claude assisted in generating ideas for visualizations, statistical analyses, and identifying patterns and relationships within the data. Different techniques like Sequential Feature Selection, Correlation Co-efficient, and Feature importance were included in implementation as well.

3. Feature Engineering:
   Claude helped me explore Ordinal encoding as well as Label Encoding for variables.

4. Claude also suggested different methods like Ensemble and Stacking to be tried but the results did not turn out so well with those methods.

The integration of generative AI in this project demonstrates the potential of AI-assisted solutions, yet personal learning and understanding of concepts was needed to tell the GenAI assistant exactly what I wanted to implement. Industry knowledge relevant to Sales would benefit in tackling complex business problems.

THE CONCLUSION

In this project, a sales forecasting model was developed using a Random Forest Regressor to accurately predict quarterly sales for a steel manufacturer's customers. The model leveraged customer-specific attributes, financial ratios, and economic indicators to capture the complex relationships influencing sales. Through data preprocessing, exploratory data analysis, and feature engineering, the data was prepared for modeling. The model was trained and evaluated using a train-test split, achieving a validation MAE of 294.45.

The sales forecasting model has the potential to provide valuable insights and support decision-making processes for the company. By accurately predicting sales, the company can optimize inventory management, resource allocation, and planning, leading to improved efficiency and profitability. The model's predictions can be used to identify potential opportunities or risks, enabling proactive measures to be taken.

Future enhancements to the model could include incorporating additional data sources, exploring advanced feature engineering techniques, and fine-tuning the model's hyperparameters to further improve its performance. Regular monitoring and updating of the model are recommended to ensure its continued accuracy and relevance in the face of changing market conditions and business dynamics.

"Without data, you're just another person with an opinion."

- W. Edwards Deming, **statistician**, and management consultant