**Car Sales Price Forecasting – Full-Stack Machine Learning Solution**

This project presents a comprehensive, end-to-end solution for forecasting car sales prices by leveraging advanced machine learning algorithms and in-depth data analysis. The core objective is to empower car dealerships and automotive businesses with data-driven insights that enhance pricing strategies, optimize inventory, and inform long-term strategic planning.

**Project Overview**

The analysis was conducted on a rich dataset comprising over **23,000** historical records of second-hand car sales. The dataset included a wide variety of features such as **buyer demographics**, **vehicle specifications**, **transaction details**, and **temporal factors**.

A robust data preprocessing pipeline was implemented:

* **Outlier removal**
* **Missing value imputation**
* **Feature engineering**  
  These steps were crucial in uncovering hidden trends and enabling accurate modeling.

**Feature Engineering & Data Enrichment**

Several high-impact features were engineered to improve model performance and interpretability:

* **Price-to-Income Ratio**: Measures affordability and purchasing power. Found to strongly influence sale prices.
* **Income Bracket**: Categorizes buyers into quartiles to enhance segmentation.
* **Company Strength**: Captures brand effect via average price per manufacturer.
* **Temporal Features**: Year, month, weekday, and a custom-designed **Seasonal Price Index** to model time-dependent effects.
* **Composite Features**: For example, PI\_plus\_model merges affordability with model characteristics.

**Correlation and mutual information analysis** confirmed that **income level**, **temporal patterns**, and **brand strength** are the top predictors of car prices.

**Modeling Strategy**

The modeling phase involved evaluating multiple machine learning models:

* **Baseline Models**: Linear Regression, Decision Tree
* **Ensemble Models**: Random Forest, Gradient Boosting, XGBoost

**Validation Strategy**:

* 80/20 train-test split
* 5-fold cross-validation
* Hyperparameter tuning via **GridSearchCV**
* Evaluation based on **Mean Absolute Error (MAE)** and **R² score**

**Performance & Model Selection**

The **Random Forest Regressor** outperformed all other models:

* **R² Score ≈ 0.997**
* **MAE ≈ 0.015**
* Excellent generalization with low bias

Feature importance analysis revealed that engineered **temporal** and **economic** features had the highest predictive value.

**Business Insights**

* **Seasonality:** Sales and prices peak in **summer** and **Q4**, helping businesses plan inventory and promotions.
* **Customer Segmentation:** Higher income segments purchase higher-value vehicles—ideal for targeted marketing.
* **Time & Location Trends:** Identifying high-performing sales periods and regions aids in resource allocation and staffing.

**Deployment & MLOps**

 **Data ingestion & cleaning**: Loaded your “Car\_sales\_Cleand.csv,” aggregated duplicate dates, and cached it for performance.

 **Multiple forecast models**: Wired up Prophet, XGBoost, ARIMA/SARIMA (with auto‑ARIMA fallback) into a unified “Forecast” page—complete with interactive Plotly charts, confidence intervals, and business metrics.

 **Model management**: Implemented a “Retrain Models” page that ingests newly uploaded forecast results, merges them into the historical dataset, and retrains Prophet, XGBoost, and ARIMA models—saving updated .pkl files.

 **Real‑time prediction**: Added a “Car Price Prediction” page that collects user‑specified features on the main canvas (not sidebar), one‑hot‑encodes them to match the Random Forest training schema, and outputs a single‑car price estimate.

 **Dev workflow**: Learned to manage Git history—resetting commits, reverting safely, pulling the latest remote version—and prepared your repo for live deployment on Streamlit Community Cloud via a requirements.txt and GitHub integration.

The final model was tracked, versioned, and deployed using **MLflow**, ensuring:

* **Reproducibility**
* **Scalability**
* **Easy integration** with business systems

This setup allows for **real-time forecasting** and **continuous model monitoring**.

**Conclusion**

This project showcases a powerful, full-stack machine learning pipeline that blends **statistical rigor**, **technical excellence**, and **business impact**. The resulting system equips stakeholders with precise predictions and actionable insights to **outperform competitors in the dynamic automotive market**.