AUTOMOTIVE PRICE PREDICTION USING ENSEMBLE LEARNING

Abstract

An Ensemble Learning Approach to Valuation: Analyzing Market Volatility and Local Feature Engineering using Random Forest Regression.

Ahmed Gul [B24F0022DS027]
Ahmed Obaid Raza [B24F0179DS015]

Contents

Project Topic and Problem Statement

Project Title

“Automotive Price Prediction using Ensemble Learning”

The used car market in Pakistan is characterized by rapid price fluctuations, long showroom wait times, and heavy reliance on informal estimations. This volatility creates significant distrust and negotiation inefficiencies for buyers and sellers.

Problem Statement: To develop a robust, objective, and data-driven system capable of accurately predicting the fair market value of used cars by quantifying the impact of technical specifications and specific, localized market dynamics (such as registration city and currency devaluation).

Project Strategy

Methodology and Scope

Our strategy was to move beyond simple averages and build an intelligent system using **Ensemble Machine Learning**. The project scope aimed to validate the market's pricing rules against the data and build a model that is both highly accurate and resilient to market noise.

Our workflow involved a structured, phased approach:

1. **Data Integrity:** Rigorous cleaning to correct errors and validate market assumptions.
2. **Exploratory Analysis:** Visually confirming market hypotheses (e.g., the Automatic Premium).
3. **Advanced Modeling:** Employing the **Random Forest Regressor** to master the complex, non-linear pricing rules that traditional models fail to capture.
4. **Model Validation:** Comparing performance against simpler baselines to prove the necessity of the complex model.

Dataset and its Attributes

The project utilized a comprehensive dataset of approximately **77,000 used car listings** sourced from Pak wheels (Kaggle).

The core data consists of 12 primary features:

Attribute	**Type**	**Market Relevance**
Price (Target)	Numerical	The value the model must learn to predict.
Year	Numerical	Primary driver of depreciation and inflation.
Engine	Numerical	Determines power, fuel efficiency, and tax bracket.
Mileage	Numerical	Direct measure of wear and tear.
Make/Model	Textual	Reflects brand loyalty, resale liquidity, and segment (e.g., Toyota vs. Suzuki).
Registered	Textual	**Crucial:** City of license plate (reflects perceived maintenance)
Transmission	Textual	Convenience factor (Automatic cars command a premium).
Assembly	Textual	Origin (Local vs. Imported, reflecting perceived quality).
Body	Textual	Shape type of vehicle (SUVs are more expensive)
Fuel	Textual	Type of fuel used (Electric are more expensive)
Color	Textual	Color of vehicles (White/ Black have more resale value)

Outline of the Project

The project followed a multi-phase design, ensuring that each step builds a reliable foundation for the next:

Phase A: Data Refinement and Custom Preprocessing

The model's success depended on addressing underlying market problems that traditional methods just can't handle:

- **Handling Missing Values (Target Price):** The few missing values in the crucial price attribute were addressed using **Hierarchical Median Imputation**. This ensured that the missing price for a car was estimated using the median price of a car that was an *exact match* (same Make, Model, and Year), providing a highly reliable market estimate.
- **Color Consolidation:** Since the original dataset contained nearly 400 minor color variations (e.g., 'Super White II', 'Pearl White'), these were aggressively grouped into **13 primary categories** (White, Black, Silver, Grey, etc.). This reduced data noise and ensured the model learned the preference for major neutral colors rather than being confused by specific shades.
- **Outlier Correction:** Data entry errors (e.g., 15,000cc typos) were corrected using **informed imputation** based on the median value of similar vehicles.
- **EV Conversion:** We addressed the growing electric vehicle (EV) segment by developing a rule to convert battery power (kWh) into an **"Equivalent CC"** (e.g., Tesla 3500cc). This ensures the model accurately values EV performance based on the established gasoline market structure.
- **Feature Translation:** Textual features like Make and Model were converted into **Market Popularity Scores** (Frequency Encoding), and features like Fuel and Transmission were converted into simple 'switch' attributes for the model to understand.

Phase B: Exploratory Analysis (Validation)

This phase served as the validation checkpoint, where visual analysis confirmed the model's key assumptions. This included:

- Confirming the dramatic effect of the **Registration City** premium.
- Analyzing the steep **Depreciation Curve**, which visually confirms the non-linear relationship of price decay over time.
- High relation of the **Engine size** with the price of the vehicle.
- Finding the high value Brands that directly affect the price.
- Confirming the hypothesis of **Automatic** and **Imported** has premium.
- Cars with widely used **Premium Colors** such as white, and black have higher resale value.

Phase C: Model Training and Selection

- **Feature Input Definition:** All features mentioned earlier were selected as input for the model, with the deliberate exception of the **city** attribute, which was excluded due to its weak correlation and redundancy with the primary registered status.

- **Data Scaling:** Continuous features (Year, Mileage) were standardized to prevent large numerical differences from biasing the model's coefficients.

- **Model Tournament:** Benchmarking the **Linear Regression** baseline against the superior **Random Forest** Regressor to justify the final choice.

- **Hyperparameter Tuning:** Systematically testing parameters (Max Depth) to select the configuration that maximizes accuracy while prioritizing **robustness**.

Project Analysis

Analysis of Predictive Power

The model comparison demonstrated the failure of simple methods and the necessity of ensemble modeling:

- **Linear Regression (Baseline):** Achieved only 44% accuracy (R^2), proving the fundamental assumption of a straight-line price relationship is incorrect for this market.

- **Polynomial Regression:** Achieved only 66% accuracy (R^2) on degree of 2. Which was still not a robust one.

- **Random Forest Regressor:** This technique excels because it can model the non-linear, segmented rules of pricing (e.g., it understands that a 2010 car has a flat depreciation rate, while a 2022 car has a vertical appreciation rate).

Key Market Insights (The "Story")

Visual analysis of the prediction results provided clear, actionable market insights:

- **The Inflation Effect:** The depreciation curve shows that prices for recent models (2018-2022) are rising almost vertically, confirming the model captures the massive impact of currency devaluation and macroeconomic inflation on vehicle assets.

- **Highest Price Drivers:** The most influential factors determining value were confirmed to be **Year of Manufacture** and **Engine Capacity**.

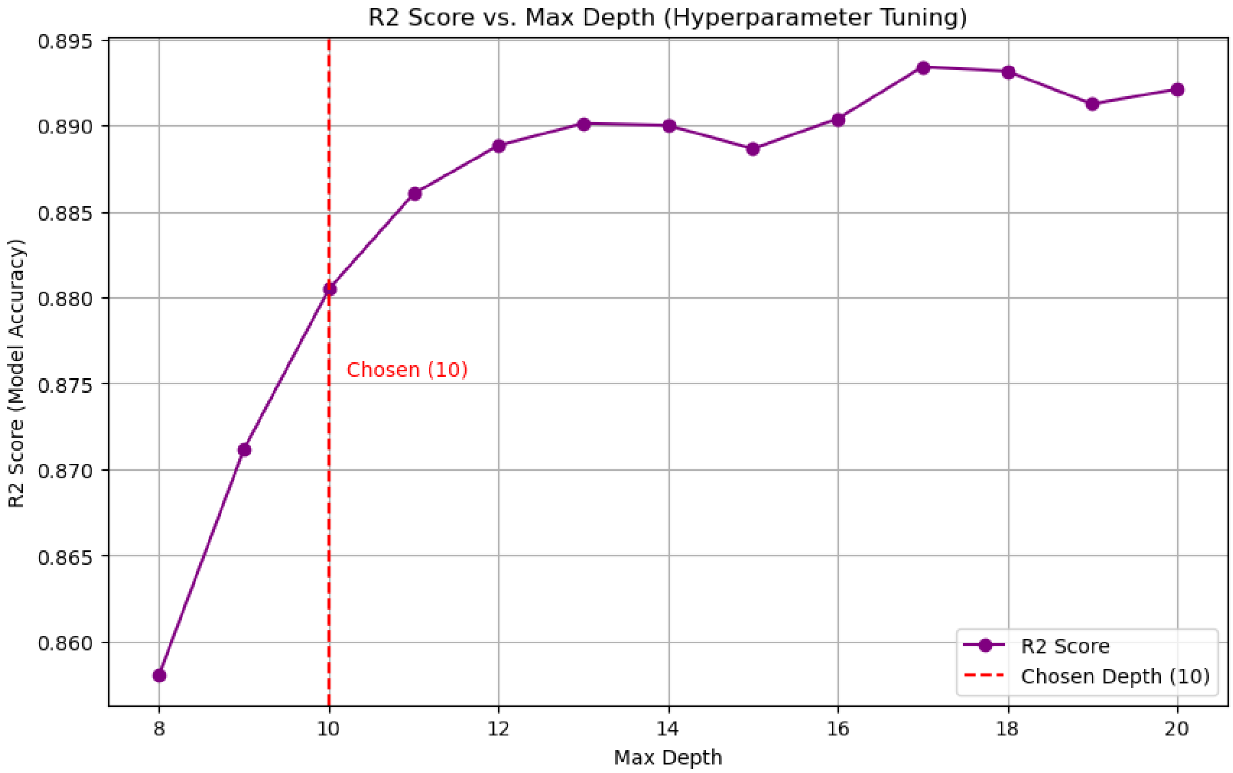
- **The On-Money Effect:** When testing the model with a brand-new Suzuki Alto, the predicted price (PKR 3.57 million) was higher than the official showroom price (PKR 3.22 million). This difference demonstrates that the model successfully captures the **Market**

Demand Premium (or "on-money") that buyers must pay to avoid long manufacturer waiting periods. This validates the model's ability to price cars based on immediate market liquidity, not just MSRP.

- **Brand Reputation**: Toyota commands premium prices due to reliability perception; luxury brands (BMW, Mercedes) significantly increase valuation.
- **High Mileage Impact:** Inverse relationship confirmed—higher mileage directly correlates with lower market value.
- **Import Premium**: Imported vehicles consistently priced higher than locally assembled equivalents, reflecting perceived quality and durability.
- **Registration City Premium**: Cars registered in major urban centers—Islamabad, Lahore, and Punjab—command significantly higher prices, reflecting maintenance perception and market demand in these regions.
- **Weak Signal and Feature Selection (Current City):** We observed that the median price for a car currently located in Lahore was negligibly different from a car currently in Islamabad or compared to other cities. This indicated that the city offers no new predictive information beyond what the model already learned from the highly correlated registered attribute (License Plate status). Given its high number of unique categories, the city attribute was dropped entirely to streamline the model and prevent unnecessary complexity and noise.

Final Model Selection

The model achieved its peak accuracy at Max Depth 17 (R^2 0.8921), but the final deployed model was deliberately set at **Max Depth 10** (R^2 0.8805). This decision was made to **regularize** the model, ensuring it focuses on general market rules rather than memorizing noisy training outliers. This guarantees the model is highly **robust** and stable when faced with new, unseen data in the real world.



Wrap Up

Overall Conclusions

The developed Random Forest pipeline is a highly effective tool for providing objective valuation in the complex Pakistani automotive market. It achieved an accuracy of **88%**, explaining most of the price variation with a reliable average error of **563,000 PKR**.

The project confirms that a strong data science solution relies heavily on:

1. **Market-Specific Feature Engineering:** Correctly quantifying local variables (EVs, Registration City, mileage etc).
2. **Ensemble Methods:** Utilizing the Random Forest to capture the complex, non-linear pricing logic established by supply, demand, and economic volatility.

Future Improvements

- **Condition Integration:** The remaining 12% error is likely driven by factors the model cannot see (e.g., physical damage). Future work should integrate image recognition or structured textual condition reports to further refine accuracy.
- **Real-time Deployment:** The final model and all preprocessing components were saved (via joblib) and are ready for deployment into a scalable, real-time web API.