Sarmaya Car: A Data-Driven Optimization Model for Used Car Investment in Pakistan

Muhammad Ayain Fida Rana & Muhammad Qasim Atiq Ullah
Department of Computer Science, Syed Babar Ali School of Science and Engineering
Lahore University of Management Sciences
Lahore, Pakistan
{25100045,25100263}@lums.edu.pk

Muhammad Rahim Subtain
Department of Management Sciences, Sulaiman Dawood School of Business
Lahore University of Management Sciences
Lahore, Pakistan
27110320@lums.edu.pk

1 Problem Statement

Owing to supply chain disruptions, rising production costs, there's been a varying trend in the sales of new cars. As per the Pakistan Automotive Manufacturers Association amidst the variation there's a decrease in net sales in the recent couple of years with the sales from 2023-2024 being 79,594 - the lowest ever since 2002-2003 (Pakistan Automotive Manufacturers Association (PAMA), 2024). At the same time we see an influx of transactions for used cars. An article on Tribune suggests that used cars account for over 60% of total vehicle transactions in Pakistan (The Express Tribune, 2024). Given the unique dynamics of the automobile industry one can also view cars as an investment asset. This is similar to buying gold or USDs where individuals buy used cars with the intention of reselling them for profit after some time. However, choosing the right car to invest in is challenging, as people usually cannot take depreciation rates, market trends, and brand retention into account.

Sarmaya[1] aims to provide buyers and sellers with details on:

- Identifying specific makes and models that investors should purchase to maximize returns while managing risk.
- Predicting expected future resale value of different vehicles based on make, model, year, and other features of interest.
- Optimal selection of a car given an initial investment upon user preferences.

2 Dataset

2.1 Dataset Collection

Instead of using pre-existing datasets, we chose to extract car listings data from PakWheels.com. To achieve this, we wrote a script using Selenium in Python, which takes a vehicle make-model pair (e.g., Toyota Corolla, Honda Civic, etc.), and then fetches all the URLs from the current page and does so iteratively until the last page.

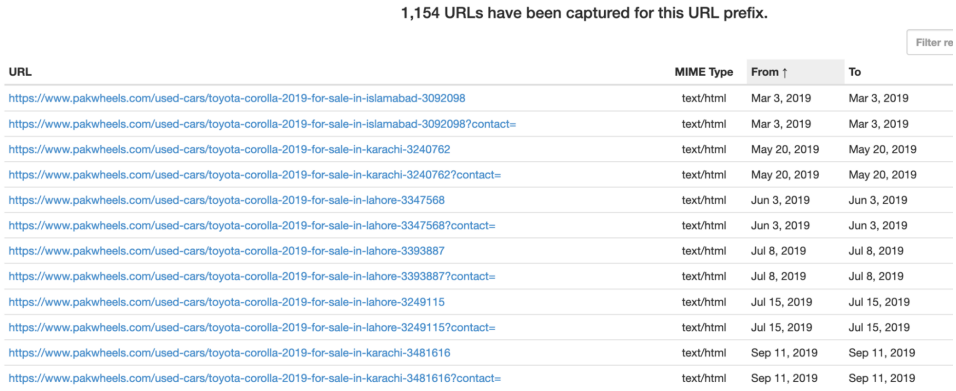
To build the dataset, we then used BeautifulSoup to scrape individual car ads from PakWheels.com. Each ad was processed by extracting relevant details and organizing them into structured JSON objects. An example JSON object is shown in Figure 1.

Once all these ad listings were scraped, their corresponding JSON objects were combined into a single dataset, stored as a CSV file. This dataset had approximately 32,000 cars listings,

[1]The final MOLP, along with the codebase and datasets, is available online.

```
{
    "Ad Ref": "9307635",
    "url": "https://www.pakwheels.com/used-cars/mg-hs-2021...",
    "Featured": 1,
    "Vehicle": "MG HS Trophy 2021",
    "Location": "Bahria Town, Lahore Punjab",
    "Model": "2021",
    "Vehicle Type": "Crossover",
    "Mileage": "31,500 km",
    "Engine Type": "Petrol",
    "Transmission": "Automatic",
    "Features": ["ABS", "AM/FM Radio", "Air Bags"],
    "Details": {"Registered In": "Punjab", "Color": "Pearl White Metallic"},
    "Price": "PKR 66.5 lacs",
    "Seller Details": "Syed Naqvi\nMember Since Oct 22, 2020",
    "Seller's Comments": "MG HS 2021 model registered in 2022 Imported..."
}
```


Figure 1: An example JSON Object



1,154 URLs have been captured for this URL prefix.

Filter re

URL	MIME Type	From ↑	To
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-islamabad-3092098	text/html	Mar 3, 2019	Mar 3, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-islamabad-3092098?contact=	text/html	Mar 3, 2019	Mar 3, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-karachi-3240762	text/html	May 20, 2019	May 20, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-karachi-3240762?contact=	text/html	May 20, 2019	May 20, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3347568	text/html	Jun 3, 2019	Jun 3, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3347568?contact=	text/html	Jun 3, 2019	Jun 3, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3393887	text/html	Jul 8, 2019	Jul 8, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3393887?contact=	text/html	Jul 8, 2019	Jul 8, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3249115	text/html	Jul 15, 2019	Jul 15, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-lahore-3249115?contact=	text/html	Jul 15, 2019	Jul 15, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-karachi-3481616	text/html	Sep 11, 2019	Sep 11, 2019
https://www.pakwheels.com/used-cars/toyota-corolla-2019-for-sale-in-karachi-3481616?contact=	text/html	Sep 11, 2019	Sep 11, 2019

Figure 2: Figure shows the URLs captured for the Toyota Corolla 2019 by the Wayback Machine

which were scrapped on 15th March; hence, making it a cross-sectional dataset i.e., data corresponding to a single point in time. This was adequate for the first component of our project, i.e., a car pricing predictor.

However, for the car pricing forecaster, we needed a longitudinal dataset e.g., the prices of Toyota Corolla 2015 from 2015 until today. For this, we used WaybackMachine, which is a digital archive of the Internet that saves snapshots of websites at various times as shown in Figure 2. Therefore, we used the Wayback CDX API to get URLs of ad listings historically. After doing so, we again retrieved the JSON objects associated with the ad listings, and created a combined CSV dataset of historical prices of about 35,000 listings.

It was assumed that prices from the ad listings reflect true market values and treated them as selling prices, despite the likelihood that the actual selling prices were lower due to negotiation biases.

We also obtained the dataset for the Consumer Price Index (CPI) from 2015 until March 2025 (Trading Economics, 2025). To estimate holding costs associated with different types of vehicles, such as Sedans, SUVs etc, we conducted a survey asking users to report their yearly costs, including token taxes, maintenance, and fuel costs.

Ad Ref	url	Make	Model	Year	Vehicle	Location	Mileage	Engine Type	Transmission	Features	Details	Price
9131958.0	https://www.pakwheels.com/used-cars/mg-hs-2021...	MG	HS	2021.0	MG HS Trophy 2021	Jinnah Town, Faisalabad Punjab	56,000 km	Petrol	Automatic	['ABS', 'AM/FM Radio', 'Air Bags', 'Air Condit...	{'Registered In': 'Punjab', 'Color': 'Pearl Wh...	PKR 61.5 lacs\n\nFinancing starts at PKR 94,49...
5236683.0	https://www.pakwheels.com/used-cars/mg-hs-2021...	MG	HS	2021.0	MG HS 1.5 Turbo 2021	Gujranwala Punjab	10 km	Petrol	Automatic	['ABS', 'AM/FM Radio', 'Air Bags', 'Air Condit...	{'Registered In': 'Un-Registered', 'Color': 'W...	PKR 59 lacs
5023147.0	https://www.pakwheels.com/used-cars/mg-hs-2021...	MG	HS	2021.0	MG HS 1.5 Turbo 2021	Faisalabad Punjab	12 km	Petrol	Automatic	['ABS', 'AM/FM Radio', 'Air Bags', 'Air Condit...	{'Registered In': 'Un-Registered', 'Color': 'B...	PKR 61 lacs\n\nFinancing starts at PKR 93,724/...

Figure 3: Snapshot of the ad listings dataset before pre-processing

2.2 Dataset Overview

A short glimpse of the original ad listings dataset is shown in Figure 3:

Some of the features obtained from the dataset are listed below:

- **Make**: The manufacturer of the car (e.g., Toyota, Honda)
- **Model**: The manufacturer of the car (e.g., Corolla, Sonata)
- **Year**: The manufacturing year of the car (e.g., 2023)
- **Mileage**: The distance the vehicle has traveled (in km)
- **Transmission**: Gear transmission of the car (Automatic/Manual)
- **Engine Type**: The fuel type of the car (Petrol, Diesel, etc.)
- **Price**: The selling price of the car listed in PKR
- **Location**: The city and province where the vehicle is listed
- **Features**: The features of the car (e.g. ABS, Air Bags, FM/AM, etc.).

A short glimpse of the original CPI dataset is shown in Figure 4.

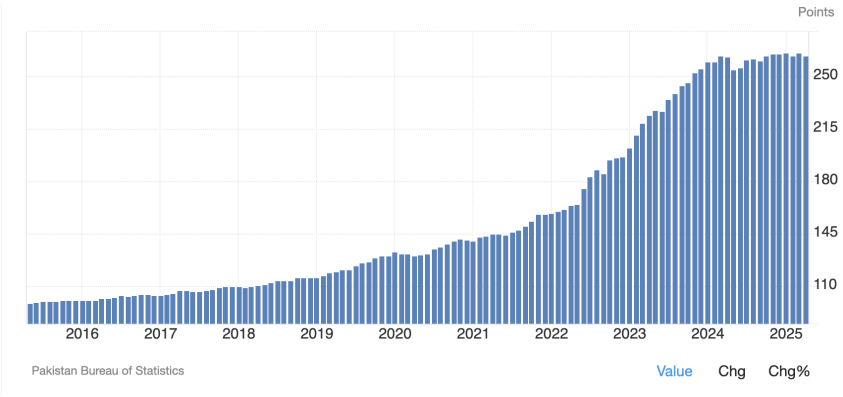


Figure 4: CPI in Pakistan averaged 99.06 points from 2001 until 2025, reaching an all-time high of 266.29 points in March 2025 and a record low of 31.12 points in July 2001 (Trading Economics, 2025)

	Make	Model	Year	Mileage	Engine Type	Transmission	Price	City	Province	Engine Capacity	...	Steering Switches	DVD Player	Navigation System	Air Bags	Climate Control	Front Speakers
0	MG	HS	2023	70000	Petrol	Automatic	7000000	Karachi	Sindh	1500	...	0	0	1	1	0	0
1	MG	HS	2023	34000	Petrol	Automatic	7000000	Lahore	Punjab	1500	...	0	0	1	1	0	0
2	MG	HS	2021	48000	Petrol	Automatic	6500000	Islamabad	Islamabad	1500	...	0	0	1	1	0	0
3	MG	HS	2021	46	Petrol	Automatic	6150000	Karachi	Sindh	1500	...	0	0	0	0	0	0
4	MG	HS	2025	30	Hybrid	Automatic	10200000	Lahore	Punjab	1500	...	0	0	1	1	0	0

Figure 5: Snapshot of ad listings dataset after cleaning and pre-processing

2.3 Dataset Cleaning and Pre-Processing

Before structuring the dataset, we performed several cleaning steps to standardize and preprocess the data for analysis:

- The **Price** column initially contained string values such as “PKR 66.5 lacs” or “PKR 7.35 crore”. We removed the non-numeric characters and mapped "lacs" and "crore" to their numerical representations.
- The **Mileage** column contained strings like “31,500 km”. We removed "," and "km" from the values to ensure they’re numerical.
- The **Location** column originally contained detailed addresses such as “PWD, Rawalpindi, Punjab”. We simplified this to extract only the city and province, and store them into new columns.

A short glimpse of dataset post-processing is shown in Figure 5.

Based on our survey results, we narrowed down the makes and models to include in our final MOLP, as PakWheels has approximately 750 unique make-model pairs, which would translate to around 8,000 decision variables if all make-model-year pairs from 2015 to 2024 were considered. Based on ad listings and user reviews data availability, survey results, and our intuition, we limited our model to 31 make-model pairs and accounted for all years from 2015 to 2024, with exceptions for imported cars or those introduced later in the timeline.

2.4 Data Analysis

To understand how numerical variables relate with each other, we analyzed the correlation matrix, as shown in Figure 6, which provides valuable insights into the relationships between numerical variables in the dataset.

- **Age and Price**: A moderate negative correlation (-0.29) exists between the age of manufacture and the price, indicating that as cars age their resale value generally drops.
- **Mileage and Price**: A weak negative correlation (-0.20) is observed between mileage and price. This suggests that vehicles with higher mileage tend to have slightly lower prices, but not as strong as age.
- **Age and Mileage**: A moderate positive correlation (0.46) is observed between age and mileage. This makes sense as older cars are logically to be driven more.

Further insights based on data exploration are discussed in Appendix A.1.

3 Price Predictive Modelling

Here, we sought to predict car resale prices based on the features, without taking into account time-based forecasting; that is, it’s based on ad listings from March 15th.

This predictive modelling is done in the following steps:

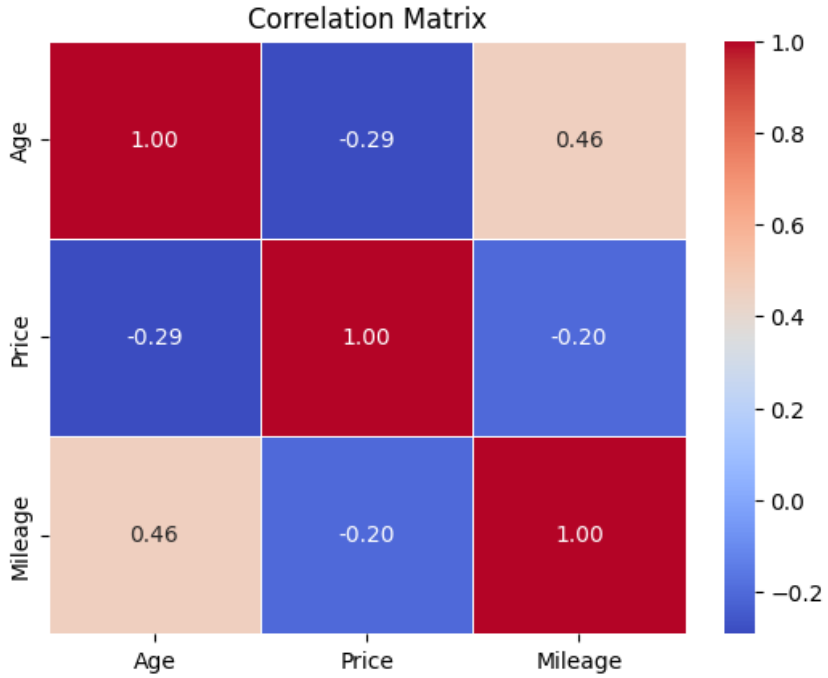


Figure 6: Correlation matrix of key numerical features

1. Prepare the dataset for the model (e.g., one-hot encoding for features like Province) and select relevant features.
2. Split the dataset into an 80:20 ratio for training and testing purposes.
3. Fit the training dataset to the models (Multiple Linear Regression and Decision Tree)
4. Evaluate the performance of the models
5. After evaluating the results, we choose the best algorithm to build the model for predicting car resale values.

3.1 Linear Regression

This model is one of the most commonly used machine learning algorithms for prediction. It predicts the dependent variable—price, in our case—based on available predictors like mileage and engine capacity. Since we have multiple independent variables, we'll use a multiple linear regression model (MLR), which can be expressed as:

$$f(X_1, X_2, ..., X_n) = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

3.1.1 Preparing the Dataset and Feature Selection

Categorical features, including '**Make**', '**Model**' '**Engine Type**', '**Transmission**', '**Province**', '**Assembly**' and '**Vehicle Type**', were converted into numerical representations using one-hot encoding.

For example, the '**Transmission**' feature, which contains either 'Automatic' or 'Manual', was replaced by a new column '**Transmission_Automatic**', where the value is set to 1 if the transmission is automatic and 0 if it is manual. Note that we do not create a separate dummy variable for 'Transmission_Manual', as Linear Regression requires that predictors are not perfectly correlated (Analytics Vidhya, 2025).

Figure 7: Distribution of raw prices across our dataset

This step was crucial because the regression model requires numerical inputs to process the data.

After the encoding, we had a significant increase in the number of features. We choose not to include the Model encodings, as doing so could reduce the generalizability of the price predictor. This is because including highly specific model indicators risked overfitting the training data, which could lead to high variance and reduce its performance on unseen data.

We then plot the distribution of prices in our dataset. As shown in Figure 7, the prices are highly right-skewed, with a long tail of very high values. These extreme outliers include cars like the Toyota Land Cruiser or Prado, and removing them would not be justifiable, as our price predictor would then lack generalizability.

To avoid this, we applied a log-10 transformation (West, 2022) to the prices, which compresses large values, makes data more systematic and reduces price variance for high-end luxury vehicles, which is evident in Figure 8.

3.1.2 Model Training and Results

We then fit the Multiple Linear Regression model onto the training dataset. The results are as shown in the table below:

METRIC	VALUE
MSE (Test)	0.00675
MSE (Train)	0.00700
R^2	0.928

Table 1: MLR Evaluation Metrics

- The **high** R^2 value indicates that approximately 92.8% of the variance in the log-transformed car prices is explained by the model.

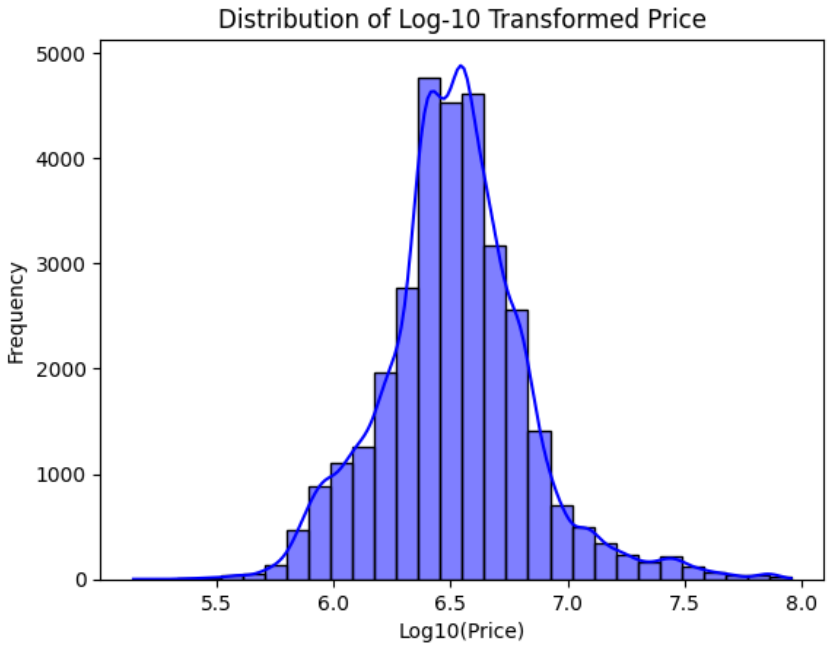


Figure 8: Distribution of log-transformed prices across our dataset

- The low MSE suggests that the model's average prediction errors are small and that it fits the data well.
- Relatively similar training and testing MSE confirms that the model is not overfitting to the data and is generalizable.

One thing to note here is that considering raw prices, we had an R^2 of 66%, which indicated a relatively weaker fit due to the higher variance in car prices. This also violated the assumption of constant variation of residuals as cheaper cars have a narrower range of price variation, while the high-end luxury ones have a much wider range.

3.1.3 Analysis

Now that we have run the model and obtained our results, let's examine what the coefficients of these predictors tells us.

In the log-linear model, coefficients translate into percentage changes, which are more meaningful in real-world decision-making.

Numerical Features As our model is on the log-10 price, a unit increase in a feature changes the log-10 price by the coefficient b.

To find the change in raw price with one unit change in a numerical feature, here's a brief overview, with detailed derivation discussed in Appendix A.3:

$$log_{10}(Price) = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

$$\frac{d}{dX_1} log_{10}(Price) = \frac{d}{dX_1}[b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n]$$

$$\frac{1}{Price \times ln(10)} \times \frac{dPrice}{dX_1} = b_1$$

$$\frac{dPrice}{dX_1} = b_1 \times ln(10) \times Price$$

Thus, a one-unit increase in a numerical variable causes the price to change by $b_1 \times ln(10)$.

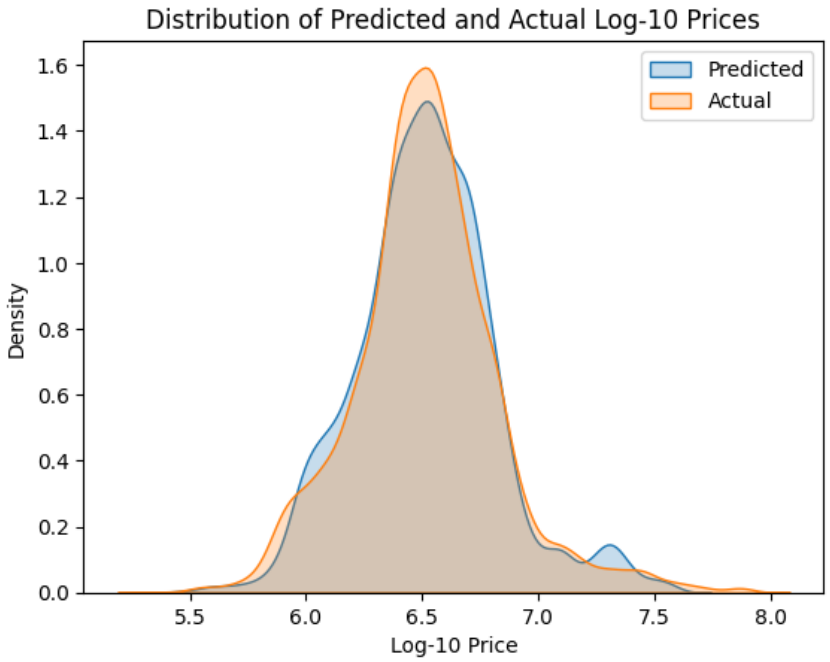


Figure 9: The plot shows that predicted log-transformed prices closely match the actual distribution. The overlap between curves indicates that the model effectively captures the full range of log-transformed price variability.

FEATURE	COEFFICIENT	INTERPRETATION
Age	-0.0244	Every additional year reduces the price by approximately 5.6%.
Mileage	-9.504×10^{-8}	Negligible per km; for a 10,000 km increase, the price reduces by approximately 0.22%.
Engine Capacity	1.0×10^{-4}	Every extra 1000cc adds approximately 23% to the price.
Transmission_Manual	-0.0469	Manual cars are priced 10.8% lower than automatic cars.
Assembly_Local	-0.1065	Local assembly lowers price by 24.5% compared to imported.

Table 2: Numerical Feature Analysis

Transmission The insights suggested by the corresponding regression coefficient makes sense because most manual transmission cars are older models, often hatchbacks, and since both predictors lower costs, this would also.

Assembly By the same token, most imported cars in our dataset are luxury vehicles, mostly SUVs, and since they incur import duties, they would cost more. Hence, their counterpart local vehicles would be priced relatively lower.

Vehicle Type Based on the regression coefficients, SUVs have the highest coefficient (relative to crossovers), followed by trucks, hatchbacks, and sedans. Thus, higher resale prices generally are in the following order:

SUV > Truck > Crossover > Hatchback > Sedan

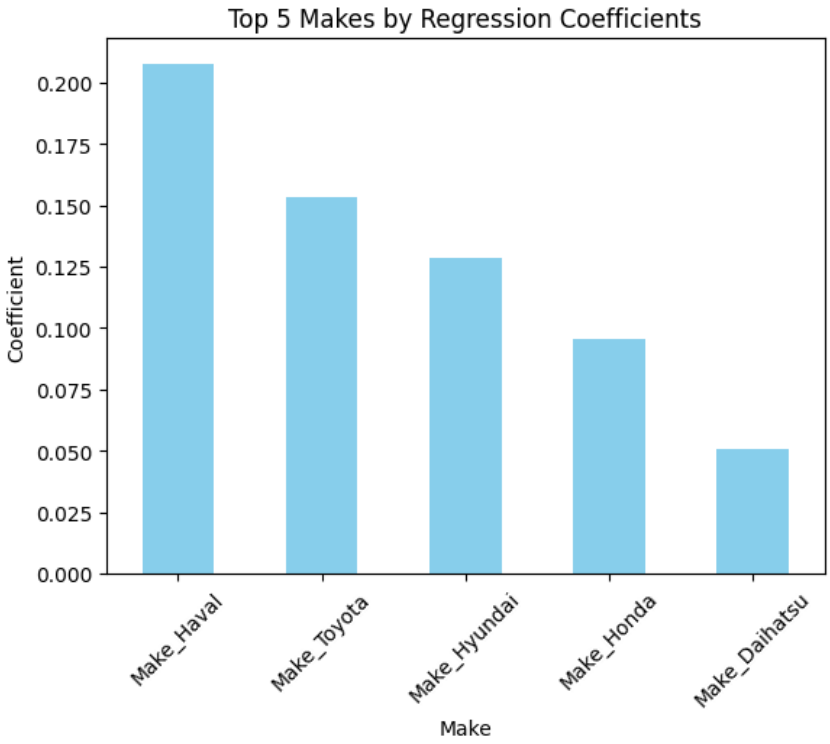


Figure 10: Top 5 makes by resale price premium (based on regression coefficients)

The relatively high price of hatchbacks is likely due to the prevalence of imported models in this category, as noted earlier.

Key Features Features such as Cruise Control, Power Windows, Power Steering, Rear AC Vents, and Heated Seats contribute approximately 8–9% to the resale value.

This is expected, as most luxury vehicles, particularly SUVs, typically include these features as standard

Engine Type Based on the regression coefficients for Hybrid (0.0786) and Petrol (0.0398), hybrid vehicles drive a stronger resale premium.

Make According to the regression coefficients, Haval commands the highest price premium, around 20.8% higher than Changan (baseline), followed by Toyota and Honda, as shown in Figure 10.

These values highlight the premium features associated with Haval's crossovers, which command prices near Rs. 1 crore, significantly higher than the baseline Changan. In contrast, Toyota and Honda, known for their reliable sedans, also have a substantial impact on prices.

Toyota's diverse lineup, including luxury SUVs and C-segment sedans and hatchbacks results in relatively lower pricing. However, when considering specific features like engine size or assembly line, Toyota's influence on price becomes particularly pronounced.

Location Comparing the regression coefficients across the densely populated regions: Islamabad, Punjab, and Sindh, we find that Sindh suffers the highest reduction in resale values, followed by Punjab and then Islamabad.

This explains why people often register cars in Islamabad, just to get a better resale value than average.

3.2 Decision Tree

We saw that linear regression required extensive data pre-processing and standardization prior to model training. We even applied log-10 transformation to prices to mitigate the impact of outliers. In contrast, decision trees effectively model non-linear relationships between features and the target variable, and are also not sensitive to skewed distributions. They can infer specific car models using attribute combinations; for example, a Toyota SUV corresponds to models like Prado, Fortuner, or Land Cruiser, which have higher prices than other Toyota cars.

Decision trees also inherently capture feature interactions. For instance, they can identify how a vehicle's type affects its value differently across regions with varying terrains. A sedan might be priced higher in Islamabad than in Sindh due to smoother roads, while SUVs could command higher prices in Sindh compared to Islamabad.

Most importantly, decision trees offer simplicity and interpretability (Mienye & Jere, 2024). They provide clear visualizations of decision-making rules, making them accessible to **non-technical stakeholders**. Each tree split represents a decision based on feature thresholds, highlighting the importance of various features in pricing.

3.2.1 Preparing the Dataset and Feature Selection

We don't need to further process the dataset obtained from the data cleaning phase initially. The only change this time will be to drop features like Cruise Control, Power Windows, etc. because of decision trees' ability to inherently capture feature interactions. We only need to encode categorical features via label encoding as the model doesn't allow string input.

3.2.2 Model Training and Results

We first sought the best hyperparameters (i.e., max depth = 10) for the decision tree. These optimized parameters helped prevent overfitting while maximizing predictive accuracy.

The model achieved a R^2 score of 0.972, indicating that it explains over **97%** of the variance in car prices.

Overall, this model effectively captures the non-linear pricing behavior of used cars without requiring extensive pre-processing.

3.2.3 Analysis

Now that we have run the model, let's explore what drives its predictions. Decision trees rely on a series of feature-based questions, and feature importance scores reveal how much each feature influences the predictions.

The plot in Figure 12 shows the top 5 features that contribute the most to the predictions of decision tree.

- **Engine Capacity** has the most predictive power, accounting for over 60% of the model's importance. This indicates that the tree relies heavily on engine size while splitting data and estimating the price. This makes sense, given that most luxury cars in our dataset have engine sizes above 2000 cc, suggesting it's strongly correlated with prices.
- **Age** accounts for approximately 35% of the predictive power, reflecting the tendency of vehicles to depreciate over time.
- All other features (**Assembly Type, Vehicle Type, and Make**) contribute relatively less, because most of this information can already be captured by Engine Capacity.
  - For example, a car with an engine size of 4600 cc would already be inferred as a high-end imported Toyota SUV, making additional splits on brand or assembly redundant in many cases.

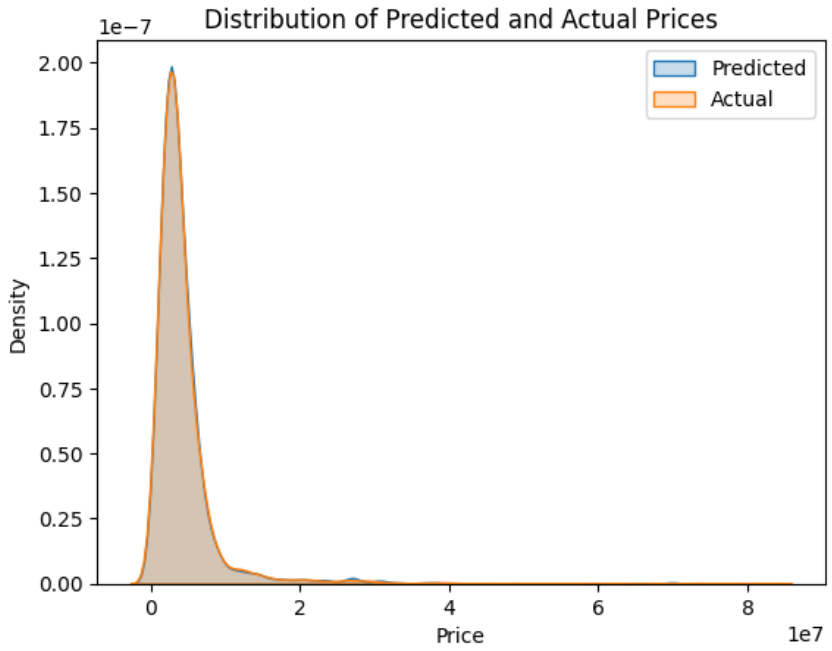


Figure 11: The plot shows that the predicted prices closely follows the actual price distribution. The near-perfect overlap between curves suggests that the model effectively captures the underlying structure of the price data, including the strong right-skew caused by the luxury cars.

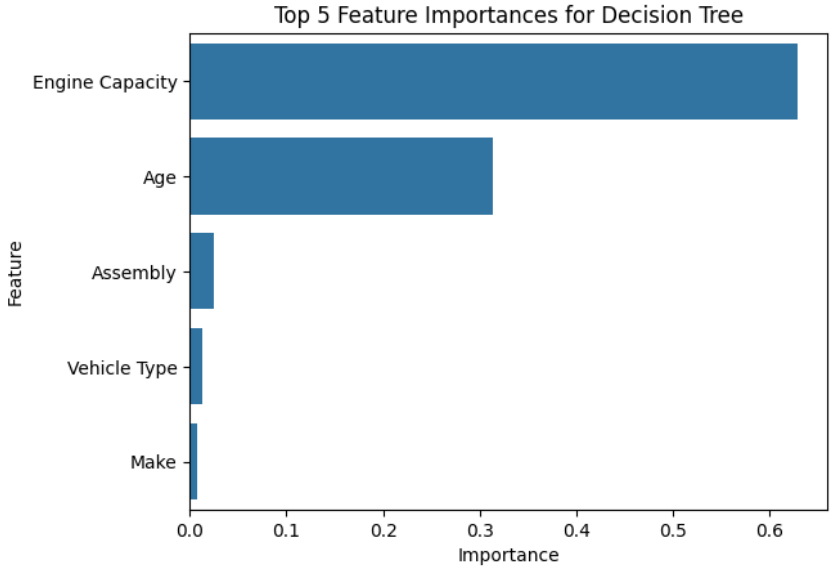


Figure 12: Top 5 features influencing the predictions

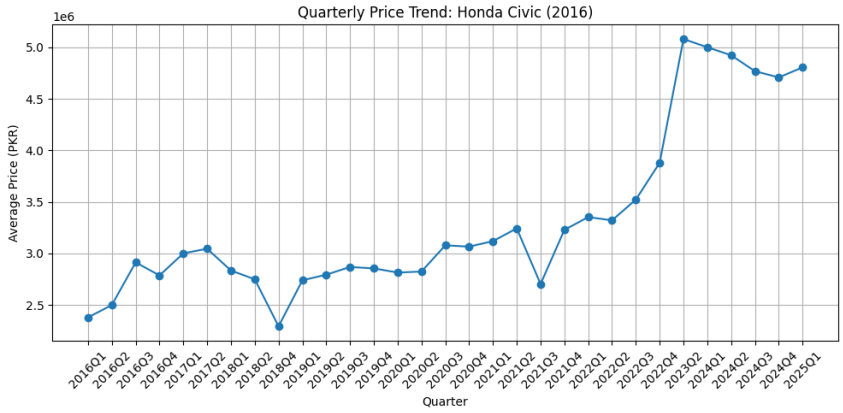


Figure 13: The plot shows that the value of the car has increased over time, but in reality the nominal prices have risen due to inflation, not the actual value of the car.

4 Price Depreciating Modeling

Here, we sought to estimate the depreciation rate at which different vehicles lose value over time, enabling more informed purchase decisions.

4.1 Methodology

- We first concatenated both the historical price listings and the present-day ad listings dataset, resulting in about 66,000 listings.
- These listings were then grouped by Make, Model, and Year and aggregated by quarterly average prices.
- To make a fair comparison between 2015's Rs. 1 and today's Rs. 1, we adjusted the quarterly average prices of each group for inflation using the Consumer Price Index (CPI) dataset.
- For each group, we calculated the quarterly depreciation rates by fitting a log-linear regression to inflation-adjusted prices.
- Finally, to estimate the depreciation per make-model pair, we computed weighted depreciation rates across time periods, with weights proportional to the number of listings in each year.

4.2 Mathematical Framework

4.2.1 Consumer Price Index

CPI tells us how expensive things have become compared to the base time. It became necessary to adjust prices; otherwise we will not be able to show the car's depreciation at all.

For instance, the average price of a 2016 Toyota Corolla in 2016-17 was around 20 lakhs, and currently it's 40 lakhs. Anyone would fall into the trap of thinking that the price doubled. In reality, this is like comparing apples with oranges.

Suppose that if the CPI was 75 in 2015 and 145 in 2018, then prices in 2018 were 1.93 times more than in 2015; similarly, Rs. 1 in 2018 was equivalent to approximately Rs. 0.5 in 2015.

Therefore, we calculate the inflation-adjusted prices as follows:

$$AdjustedPrice_t = Price_t \times \frac{CPI_{current}}{CPI_t}$$

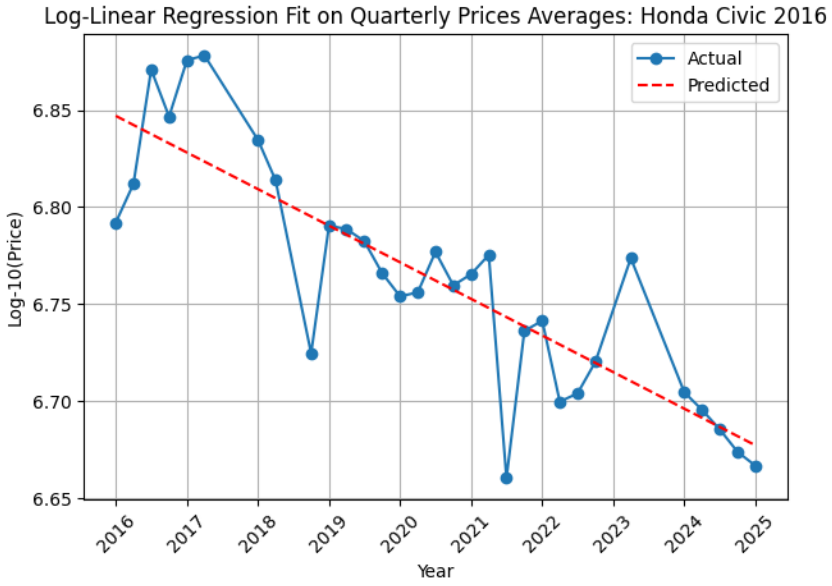


Figure 14: Inflation-adjusted log-transformed prices for Honda Civic (2016) across years

4.2.2 *Log-Linear Regression*

We used log10-transformation above for price prediction as well. Studies have shownn vehicle prices are more likely to depreciate exponentially (Ackerman, 1973) rather than linearly, log-transformed prices make sense. The relationship becomes as follows:

$$log_{10}(P_t) = log_{10}(P_0) + kt$$

Where:

- P_t is inflation-adjusted price at time t
- P_0 is inflation-adjusted starting price
- k is the slope

To find the rate of change of price over time, we take derivative wrt. t on both sides

$$\frac{d}{dt} log_{10}(P_t) = \frac{d}{dt}[log_{10}(P_0) + kt]$$

$$\frac{1}{P \times ln(10)} \times \frac{dP_t}{dt} = k$$

$$\frac{dP_t}{dt} = k \times ln(10) \times P_t$$

The slope will be negative, as prices are expected to fall over time. The quarterly depreciation rate would then be calculated as $-k \times ln(10) \times P_t \times 100\%$.

This shows that the percentage drop in price per time period remains approximately constant, which aligns well with an exponential decay model.

For instance, consider the plot in Figure 14 for the Honda Civic 2016

The equation of line is defined below:

$$log_{10}(P_t) = 6.85 - 0.005661t$$

Thus, the depreciation rate would then be $-(-0.005661 \times ln(10) \times 100) \approx 1.30\%$ per quarter.

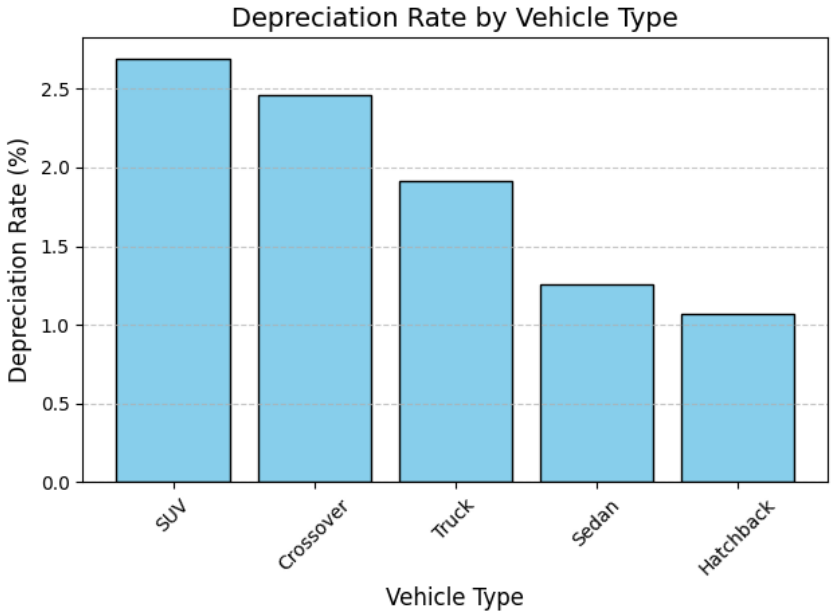


Figure 15: Quarterly average weighted depreciation rates across our included vehicle types

4.2.3 Weighted Depreciation Rates

The depreciation per make-model pair is calculated using a weighted average based on the number of listings. This smooths out year-to-year volatility and assigns a single, robust depreciation rate to a particular car.

$$\text{Weighted Depreciation Rate} = \frac{\sum_i Rate \times Listings_i}{\sum_i Listings_i}$$

One thing to note is that the latest 2025 models were not included for each make-model pair, and some other make-model-year groups with very few listings were dropped under the assumption that they introduce noise and are not representative of the true market.

4.2.4 Results and Insights

The plot in Figure 15 shows the average quarterly depreciation percentage for various vehicle types.

- SUVs and Crossovers experience the highest depreciation rates. This is likely due to their higher initial prices and the fact that these types of model mainly include imported cars in the SUV segment, which depreciate rapidly once they are registered. Most Crossovers are mainly from Changan, KIA, and Hyundai in our dataset, and they lack brand retention compared to Toyota, so they might depreciate more quickly.
- Sedans and Hatchbacks depreciate the slowest, making them relatively value-stable over time. Hatchbacks are usually priced lower initially, so they are more likely to retain their value.

The plot in Figure 16 illustrates the average quarterly depreciation rates of cars classified by their make:

- Hyundai and MG show the highest quarterly average depreciation rates. This is due to

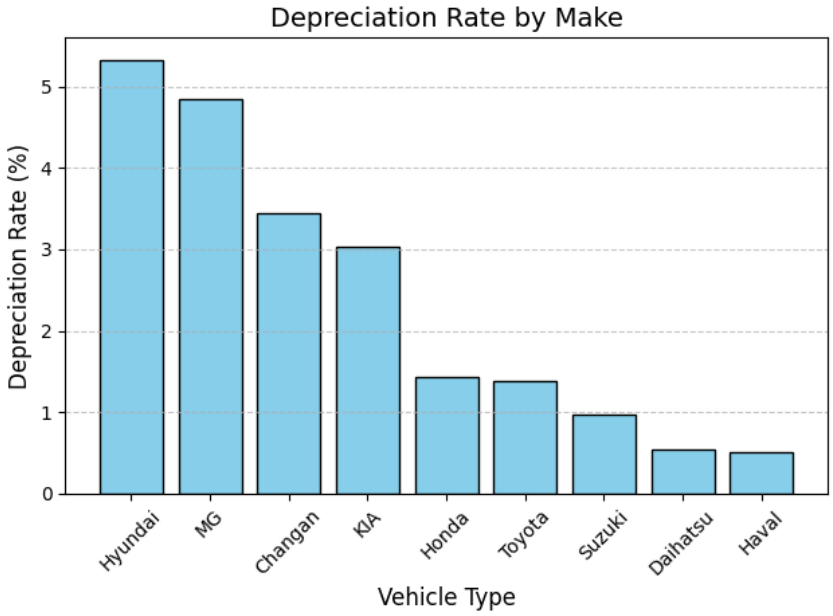


Figure 16: Quarterly average weighted depreciation rates across our included vehicle makes

  - A less established resale market in Pakistan, which is also the case with other post-COVID brands such as Changan and KIA.
  - Limited brand retention compared to Honda's and Toyota's legacy.
  - MG[2] users experienced significant dissatisfaction in the initial years due to poor fuel economy and AC cooling issues, which could explain this.
  - Hyundai launches new generations of its cars in a shorter time period than legacy brands, which could explain why older models tend to depreciate faster.
- Honda, Toyota, and Suzuki maintain moderate to low depreciation likely due to:
  - Higher demand in the used car market, ensuring value retention.
  - Strong after sales support, spare part availability, and brand trust.

5 Optimization Model

Towards the end of our project, we will build a LP model to suggest an optimal make, model, and year of a vehicle, subject to budget constraints and the buyer's perception of depreciation, value for money, holding costs, comfort, luxury, or the likelihood of reselling quickly within days.

Unlike a regular LP, we have more than one objective to optimize here, and each is different in nature, i.e., either minimizing or maximizing. Thus, we require a different form of linear program, i.e., a Multi-Objective Linear Program (MOLP), where we seek to get the best solution that meets each of the user's desired goals while minimizing undesirable weighted deviations from those goals (Lootsma, 1999).

5.1 Objectives

The goal is simple yet complex: to select the optimal used car based on multiple conflicting goals. Rather than choosing a car based on a single factor, such as depreciation, we take into account other factors like value for money, holding costs, and resale potential. Table 3 gives an overview of each of these goals.

[2]This refers to the MG HS, as it is the only model considered in our dataset

No.	OBJECTIVE	MIN/MAX	DESCRIPTION
1	Holding Costs	Minimize	Maintenance, insurance, tax — costs incurred for keeping the car
2	User Satisfaction	Maximize	Derived from review-based scores on comfort, reliability, features, etc.
3	Value Efficiency	Maximize	Derived from ratio of satisfaction to total ownership cost, and fuel efficiency
4	Depreciation	Minimize	Quarterly % decrease in price; lower means better long-term value retention
5	Popularity	Maximize	Based on number of listings in the dataset; vehicle that is likely to be most easily sold

Table 3: Goals for optimal vehicle selection

5.2 MOLP Formulation

5.2.1 Decision Variables

Let:

- x_i be if make-model-year i is selected
- y_j be if make-model j is selected
- z_k be if vehicle type k is selected
- t_l be if vehicle of year l is selected

5.2.2 Deviational Variables

Let:

- d_i^+ represent the amount by which goal i is over-achieved
- d_i^- represent the amount by which goal i is under-achieved

5.2.3 Parameters

Let:

- w_i^+ and w_i^- represents the weight associated with d_i^+ and d_i^- respectively
- P_i^+ and P_i^- represents the priority associated with d_i^+ and d_i^- respectively
- T_i be the optimal value of goal i
- B be the total budget
- p_i be the average price of make-model-year i
- s_j be the user satisfaction score for make-model-year j
- v_j be the value efficiency score for make-model-year j
- h_k be holding costs associated with vehicle type k
- u_j be the popularity score for make-model j
- r_j be the quarterly depreciation rate for make-model j

5.2.4 Objective Function

Our objective is to minimize the weighted sum of percentage deviations, along with a small penalty associated with the selected vehicle's year l.

The reason for using percentage deviations instead of absolute deviations is that the scales of each deviation are different. For instance, a deviation for the user satisfaction score goal would be between 0 and 5, while the budget constraint deviations can be in thousands, dominating its effect.

Additionally, all parameters beyond car prices are specific to the make-model or vehicle type level, which is a limitation of our model further discussed in Section 7. Once the solver selects a make-model (e.g., Toyota Corolla), it is forced to pick the latest available year within that category that satisfies the budget constraint. This is achieved by associating a small penalty with each year in ascending order, with 2015 having the maximum penalty and 2024 having the least.

This approach ensures fair weightage is given to goals as assigned by the user.

$$\text{Objective: } \textbf{MIN } \sum_{i=1}^{5} \frac{1}{T_i}[P_i^+ w_i^+ d_i^+ + P_i^- w_i^- d_i^-] + \sum_{l=2015}^{2024} t_l \times [0.1 - (l - 2015) \times 0.01]$$

5.2.5 *Soft Constraints*

Holding Costs: $\sum_k h_k z_k - d_1^+ + d_1^- = T_1$

User Satisfaction: $\sum_j s_j y_j - d_2^+ + d_2^- = T_2$

Value Efficiency: $\sum_j v_j y_j - d_3^+ + d_3^- = T_3$

Depreciation: $\sum_j r_j y_j - d_4^+ + d_4^- = T_4$

Popularity: $\sum_j u_j y_j - d_5^+ + d_5^- = T_5$

Note that T_i is determined by setting each goal as an objective function in a separate linear program, detailed in the Appendix A.6.2.

5.2.6 *Hard Constraints*

Budget Constraint: $\sum_i p_i x_i \leq B$

One Vehicle Selected Constraint: $\sum_i x_i = 1$ and $\sum_k z_k = 1$

Link Vehicle Type & Make-Model: $y_j - z_k = 0$ for each make-model j of vehicle type k

Link Make-Model-Year & Make-Model: $y_j - \sum_i x_i = 0$ for each unique vehicle i, defined by its make, model j

Link Make-Model-Year & Year: $\sum_{i \in I_l} x_i - t_l \leq 0$ for each unique year l, where I_l is the set of vehicles of year l

Binary Variable Constraint: $x_i \in \{0, 1\}$, $y_j \in \{0, 1\}$ and $z_k \in \{0, 1\}$

Non-negativity Constraint: $d_i^+ \geq 0$ and $d_i^- \geq 0$

This LP can be further modified to generate different recommendations for various buyer types by adjusting the weights or adding/removing constraints. For instance, we can set the z_k corresponding to Sedan to 1 in the constraints, so that we only consider that segment of cars.

5.2.7 *Results and Insights*

For the purposes of understanding the decision-making process, let's explore two different budget tiers, and within each we will evaluate multiple types of car buyers and their personas within each budget tier reflecting different preferences regarding vehicle holding costs, user satisfaction, value efficiency, depreciation, and resale potential. For each profile[3], we assign pre-emptive priority levels and weights within those levels, which are available in the Appendix A.6.3.

1. **PKR 1 Crore (No Hatchbacks)**

   Premium-Conscious Buyer Such a buyer prioritizes comfort, luxury, prestige, and satisfaction over cost. They prefer stable long-term ownership and are less concerned about associated operational costs and depreciation.

   Resale-Oriented Investor Such a buyer prioritizes comfort, luxury, prestige, and satisfaction over cost. They prefer stable long-term ownership and are less concerned about associated operational costs and depreciation.

2. **PKR 50 Lacs**

   Liquidity-Seeking Budget Strategist Such a buyer prefers a car that retains most of its value and is easy to sell. In short, treating the purchase as a semi-liquid asset. Most of these buyers fall into categories of individuals who have some cash in hand that they want to invest in a way that allows them to easily retrieve it when needed.

   Value-Oriented Quality Seeker Such a buyer prioritizes a balance between comfort, features, and financial sense. While value for money and user satisfaction are their top concerns, they are somewhat concerned with depreciation and resale value and are least concerned about holding costs. In short, they're willing to incur operational expenses for a better overall experience.

Results Table 4 shows the recommendations made by our model for each buyer profile.

BUYER PROFILE	BUDGET TIER	SELECTED VEHICLE
Premium-Conscious Buyer	1 Crore	Changan Ohsan X7 2024
Resale-Oriented Investor	1 Crore	Toyota Prius 2021
Liquidity-Seeking Strategist	50 Lacs	Toyota Aqua 2017
Value-Oriented Quality Seeker	50 Lacs	Toyota Corolla 2019

Table 4: Optimal vehicle recommendations

Side Quest An important thing to note here is that we obtained some cars that were not originally prescribed by either of the goals when set as an objective function. This is because we used a MINIMAX objective function, where our objective is to minimize the maximum deviation from any goal (Ahuja, 1985). Thus, our proposed model now has:

- Objective Function: **MIN** Q
- One additional constraints for each deviational variable of the form:
  $P_i \times \frac{w_i^+}{T_i} \times d_i^+ \leq Q$ and $\frac{w_i^-}{T_i} \times d_i^- \leq Q$

6 Recommendations

We have explained the recommendations and results in each of the sections above, but to consolidate and provide an overview, we can categorize them separately for sellers and buyers:

[3]These personas, along with assigned priorities, have been sourced from (CBT News, 2023) and adapted using GPT-4o and our understanding of the local market.

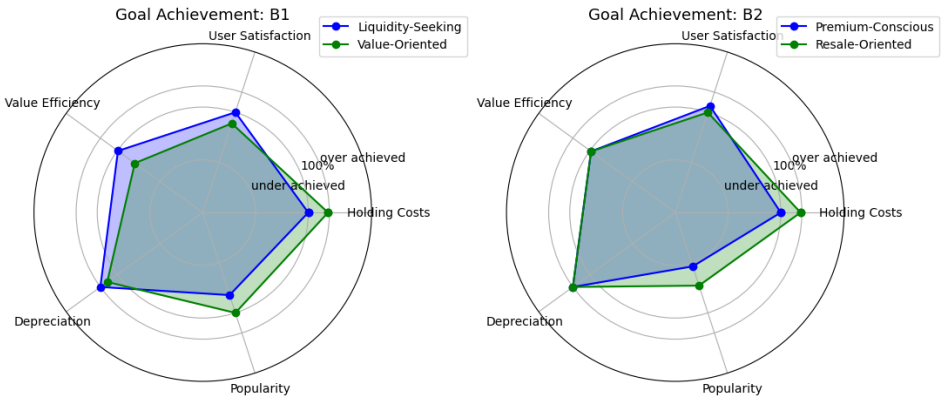


Figure 17: As per the goal-achievement radars, goals closely aligning to 100% are those with priority level 1 and greater base weights within each priority level. This demonstrates the benefit of our proposed Goal Programming model, where we can achieve the most optimal option in accordance with our priorities while minimizing deviations on each goal.

6.1 Sellers

- They can utilize decision trees and linear regression models to obtain a competitive estimate of their car's value based on mileage and other variables such as brand value, comfort, fuel efficiency etc.
- Showcasing low depreciation rates associated with their car can increase its perceived value, allowing sellers to propose long-term value.
- Estimates of fair pricing lead to more transparency and help build buyer confidence and customer retention.
- Targeted marketing can play an effective role; for example, Figure 18 shows that SUVs tend to command higher prices in Sindh compared to Islamabad, although the port is located in Karachi, so there are also no freight charges. Despite this, customers in Sindh are willing to pay a premium for the status associated with SUVs. Thus, showroom sellers can strategically acquire inventory or implement region-specific marketing.

6.2 Buyers

- Like sellers, buyers can also use price predictors to estimate the current market value of their desired vehicle and avoid overpaying. This increases the possibility of them getting a relatively better car at the same price or at a relatively lower price.
- Depreciation models can help them assess the value of their car in the coming years and enable them to make an informed decision.
- They can leverage MOLP to get personalized car recommendations aligned with their priorities for cost savings, comfort, and resale potential within their budget; thereby, saving a lot of buyers time.

7 Limitations and Future Work

To further enhance the model's performance and applicability, future work could include a more granular level of analysis, such as differentiating between variants of each model. For instance, the Honda Civic has three distinct variants: Standard, Oriel, and Turbo, each priced differently. Currently, we consider them as one, which contributes to the significant standard error in our predictions.

To improve estimates of depreciation rates, we can also incorporate macro-indicators like petrol prices and interest rates.

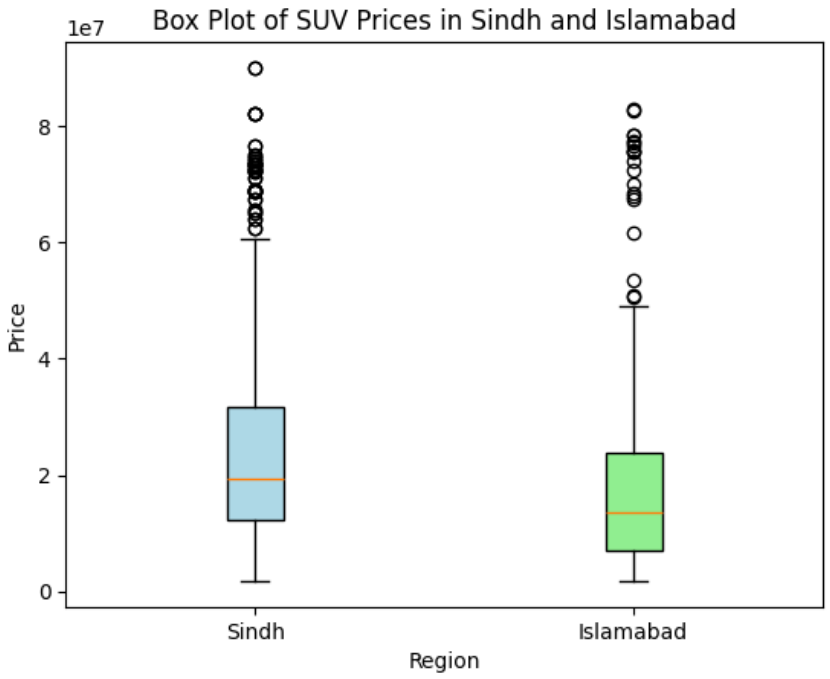


Figure 18: The plot shows the

Furthermore, our analysis suggests that newer car generations tend to depreciate at a slower rate than older ones; for example, the Corolla X depreciates less than its predecessors and is less dependent on the year of manufacture. Currently, depreciation rate parameters are associated with each make-model, which could be associated with make-model-year to address this issue.

We have assumed that used cars depreciate at an exponential rate based on (Ackerman, 1973); however, some literature suggests that while exponential curves can model depreciation, they might need additional adjustments (via auxiliary curves) to accurately fit real-world data (Carman, 1956).

As initially mentioned, our model could be further developed to suggest a portfolio of cars. This can be achieved by removing constraints limiting the selection of only one make-mode-year decision variable.

One could also incorporate sentiment analysis of user reviews as additional user-relevant classes in the MOLP and further improve our buyer profiles to be more specific in terms of user requirements.

An LLM chain[4] can be set up where user input can be utilized to automatically assign priorities and weights to relevant goals, which can then be fed to our MOLP.

8 Summary

To sum up, we have utilized 1) predictive modeling, 2) curve fitting for depreciation estimation, and 3) a multi-objective linear program. Decision trees appeared to be the best model, achieving an R^2 score of 0.972, for predicting prices even with fewer features, due to their ability to infer hidden relationships and capture non-linear trends effectively. Furthermore, fitting linear graphs to log-transformed prices effectively captured the quarterly depreciation rates for each make-model-year, which were then combined to assign a value specific to

[4]A simplified version of this is demonstrated in Appendix A.6.4.

make and model only. Finally, the MOLP helped identify a vehicle within a buyer's budget constraint and optimize according to their idea of best fit. We found that hatchbacks and sedans generally depreciate the least. For buyers who prioritize minimal depreciation and high resale potential, the Suzuki Alto is the best option, but if hatchbacks are excluded, then the Toyota Corolla is the best choice, which is not surprising. The Haval H6 seems to be the most optimal crossover choice for individuals prioritizing both maximum comfort and minimal depreciation.

Acknowledgments

We thank the entire DISC 212 teaching staff, especially Dr. Zaid Saeed Khan, and our assigned TA Rafay, for their assistance throughout this project.

References

Susan Rose Ackerman. Used cars as a depreciating asset. *Western Economic Journal*, 11(4):463, dec 1973. URL https://www.proquest.com/scholarly-journals/used-cars-as-depreciating-asset/docview/1297284331/se-2.

R.K Ahuja. Minimax linear programming problem. *Oper. Res. Lett.*, 4(3):131–134, October 1985. ISSN 0167-6377. doi: 10.1016/0167-6377(85)90017-3. URL https://doi.org/10.1016/0167-6377(85)90017-3.

Analytics Vidhya. Effect of Multicollinearity on Linear Regression, 2025. URL https://medium.com/analytics-vidhya/effect-of-multicollinearity-on-linear-regression-1cf7cfc5e8eb.

Lewis A. Carman. Non-linear depreciation. *The Accounting Review*, 31(3):454–491, 1956. ISSN 00014826. URL http://www.jstor.org/stable/242176.

CBT News. 6 Automotive Customer Types and the Lifetime Value They Bring to Car Dealers, May 2023. URL https://www.cbtnews.com/6-automotive-customer-types-and-the-lifetime-value-they-bring-to-car-dealers/.

Freerk A. Lootsma (ed.). *Multi-Objective Linear Programming*, pp. 229–257. Springer US, Boston, MA, 1999. ISBN 978-0-585-28008-0. doi: 10.1007/978-0-585-28008-0_10. URL https://doi.org/10.1007/978-0-585-28008-0_10.

Ibomoiye Domor Mienye and Nobert Jere. A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access*, 12:86716–86727, 2024. doi: 10.1109/ACCESS.2024.3416838.

Pakistan Automotive Manufacturers Association (PAMA). Historical Data 1995-2024, Nov 2024. URL https://pama.org.pk/wp-content/uploads/2024/11/Historical-Data-1995-2024-1.pdf.

The Express Tribune. Demand for used imported car thrives amid rising prices, Dec 2024. URL https://tribune.com.pk/story/2496715/demand-for-used-imported-car-thrives-amid-rising-prices.

Trading Economics. Pakistan Consumer Price Index (CPI), 2025. URL https://tradingeconomics.com/pakistan/consumer-price-index-cpi.

R. M. West. Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry*, 59(3):162–165, 2022. doi: 10.1177/00045632211050531. URL https://doi.org/10.1177/00045632211050531.

A Appendix

A.1 Exploratory Data Analysis

This section discusses several key insights revealed from the exploration of the dataset.

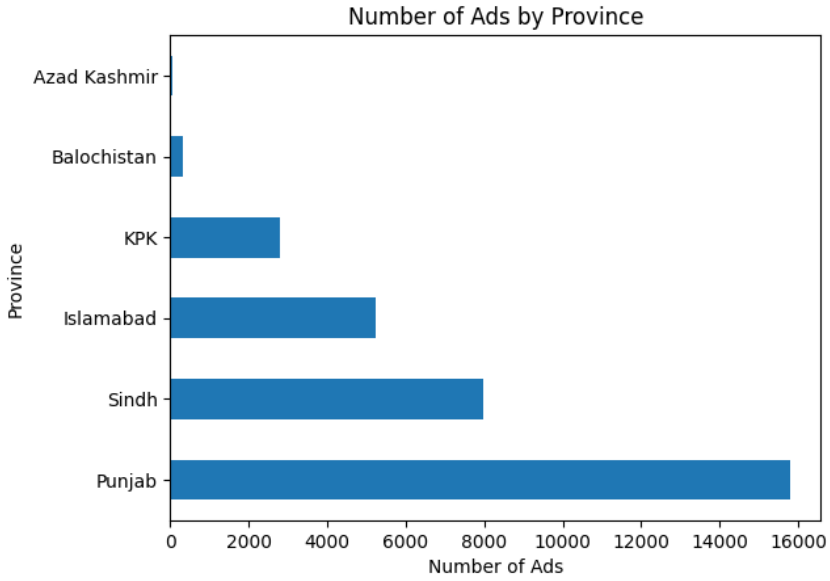


Figure 19: The Distribution of ad listings by Province gives insights related to the geographical concentration of vehicle listings across Pakistan. Punjab has the highest number of ad listings, indicating that it is the dominant region for vehicle trading and advertisements in the dataset.

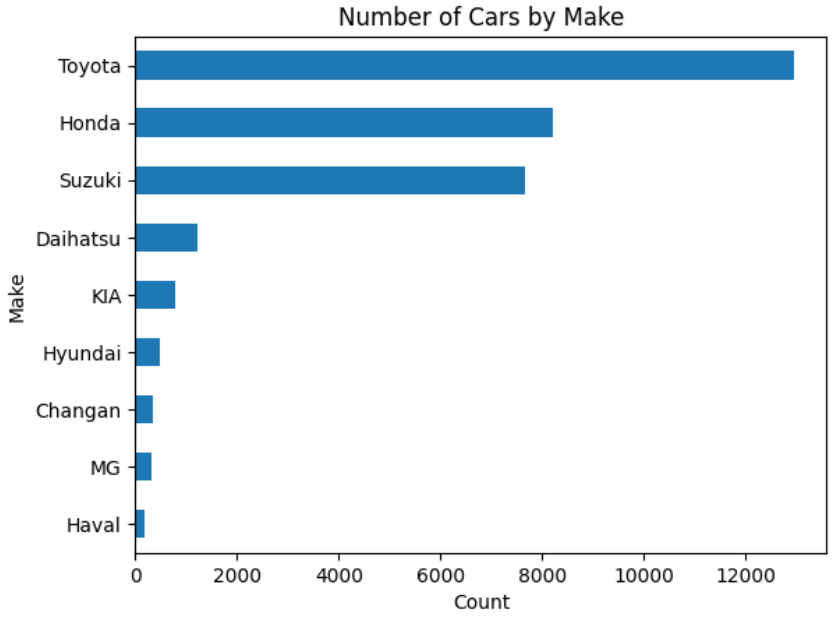


Figure 20: The bar plot shows vehicle makes popularity across our dataset. Leading the chart, Toyota, Honda, and Suzuki dominate the used car market, with Toyota standing out as the most advertised brand. This trend reflects the brand's value, be it due to reliability, affordability, or strong resale prices.

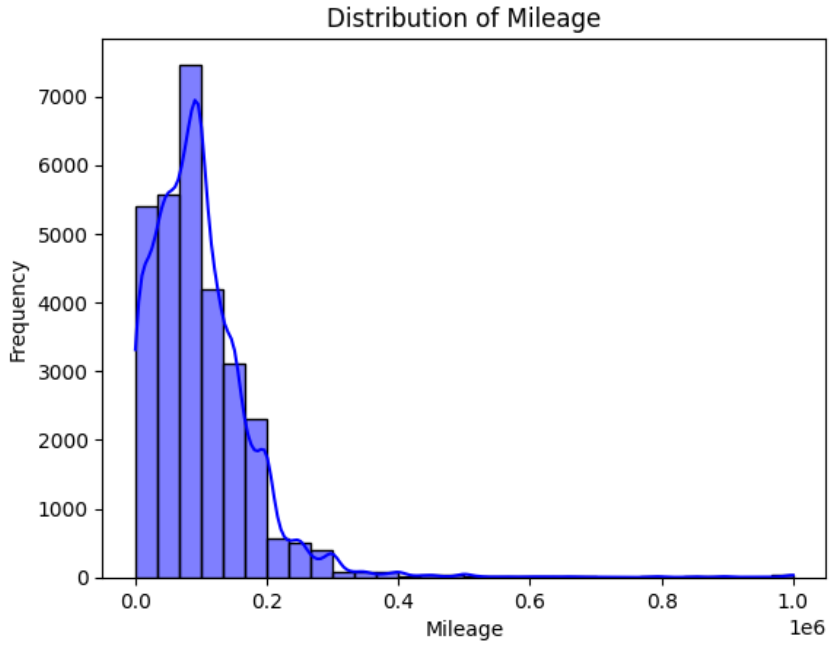


Figure 21: The mileage distribution is strongly skewed to the left, with the majority of vehicles having a mileage below 200,000 km. This trend reflects the market focus on relatively newer or moderately used vehicles, suggesting that the coefficient for mileage in the OLS Regression would be negative.

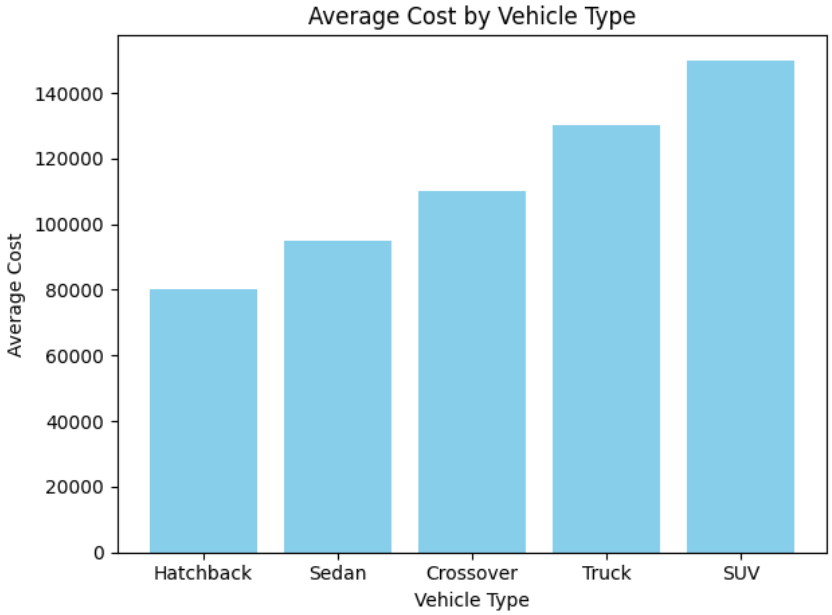


Figure 22: According to our survey, SUVs and trucks have the highest operational costs, with slight variations in crossovers and sedans due to different engine types, i.e., petrol or hybrid cars. However, what was important to us was determining the vehicle type order by increasing costs, so that we could assign appropriate penalties in the LP model moving forward.

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.928
Model:	OLS	Adj. R-squared:	0.928
Method:	Least Squares	F-statistic:	8930.
Date:	Wed, 30 Apr 2025	Prob (F-statistic):	0.00
Time:	12:43:34	Log-Likelihood:	27308.
No. Observations:	25721	AIC:	-5.454e+04
Df Residuals:	25683	BIC:	-5.423e+04
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.7599	0.015	436.364	0.000	6.730	6.790
Age	-0.0244	0.000	-213.187	0.000	-0.025	-0.024
Mileage	-9.504e-08	7.56e-09	-12.576	0.000	-1.1e-07	-8.02e-08
Engine Capacity	0.0001	1.77e-06	57.533	0.000	9.81e-05	0.000
Keyless Entry	-0.0008	0.002	-0.445	0.656	-0.005	0.003
Power Locks	-0.0346	0.003	-13.826	0.000	-0.039	-0.030
Sun Roof	0.0007	0.002	0.345	0.730	-0.003	0.004
Rear Seat Entertainment	-0.0125	0.010	-1.265	0.206	-0.032	0.007
Alloy Rims	0.0036	0.001	2.607	0.009	0.001	0.006
USB and Auxillary Cable	0.0066	0.003	1.997	0.046	0.000	0.013
Immobilizer Key	0.0148	0.002	7.332	0.000	0.011	0.019
Cruise Control	0.0382	0.002	20.045	0.000	0.034	0.042
CoolBox	0.0038	0.003	1.389	0.165	-0.002	0.009
Power Mirrors	0.0212	0.002	9.065	0.000	0.017	0.026
Rear AC Vents	0.0324	0.005	6.945	0.000	0.023	0.041
Heated Seats	0.0116	0.006	2.017	0.044	0.000	0.023
Air Bags	0.0142	0.002	7.913	0.000	0.011	0.018
Climate Control	-0.0119	0.004	-3.066	0.002	-0.019	-0.004
Front Speakers	-0.0213	0.003	-7.100	0.000	-0.027	-0.015
Power Windows	0.0282	0.003	10.081	0.000	0.023	0.034
Front Camera	0.0339	0.005	6.346	0.000	0.023	0.044
Rear Camera	0.0176	0.004	4.973	0.000	0.011	0.025
Power Steering	0.0375	0.003	14.272	0.000	0.032	0.043
Engine Type_Diesel	-0.0132	0.007	-1.822	0.068	-0.027	0.001
Engine Type_Electric	-4.77e-15	3.65e-15	-1.307	0.191	-1.19e-14	2.38e-15
Engine Type_Hybrid	0.0786	0.007	11.758	0.000	0.066	0.092
Engine Type_LPG	-0.0610	0.020	-3.025	0.002	-0.101	-0.021
Engine Type_Petrol	0.0398	0.006	6.366	0.000	0.028	0.052
Transmission_Manual	-0.0469	0.002	-28.405	0.000	-0.050	-0.044
Province_Balochistan	-0.0460	0.014	-3.199	0.001	-0.074	-0.018
Province_Islamabad	-0.0619	0.013	-4.615	0.000	-0.088	-0.036
Province_KPK	-0.0542	0.013	-4.025	0.000	-0.081	-0.028
Province_Punjab	-0.0790	0.013	-5.909	0.000	-0.105	-0.053
Province_Sindh	-0.1005	0.013	-7.495	0.000	-0.127	-0.074
Assembly_Local	-0.1065	0.002	-63.618	0.000	-0.110	-0.103
Vehicle Type_Hatchback	-0.1746	0.003	-62.072	0.000	-0.180	-0.169
Vehicle Type_SUV	0.3830	0.005	82.178	0.000	0.374	0.392
Vehicle Type_Sedan	-0.0312	0.003	-12.152	0.000	-0.036	-0.026
Vehicle Type_Truck	0.2048	0.006	33.109	0.000	0.193	0.217

Omnibus:	5094.116	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	172169.717
Skew:	0.061	Prob(JB):	0.00
Kurtosis:	15.674	Cond. No.	1.31e+16

Figure 23: Regression results summary

A.2 Linear Regression Summary

Figure 23 summarizes the results of the OLS regression model that we used to predict the log-transformed car prices.

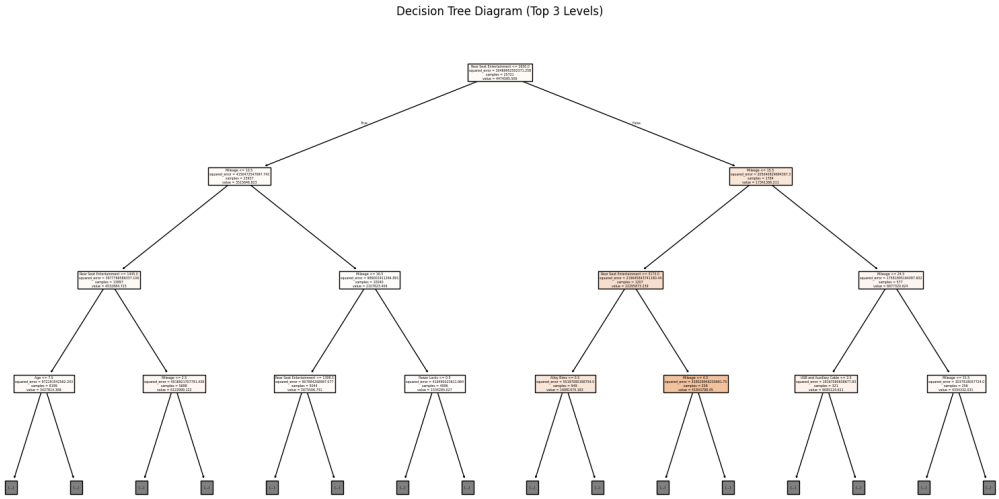


Figure 24: Decision tree visualization

A.3 Mathematical Basis of the Interpretation of Log-Linear Model

Consider a log-linear relation: $log_{10}(y) = bx$. Let's first determine the rate of change of y wrt. x.

$$log_{10}(y) = bx$$
$$\frac{d}{dx} log_{10}(y) = \frac{d}{dx}[bx]$$
$$\frac{1}{y \times ln(10)} \times \frac{dy}{dx} = b$$
$$\frac{dy}{dx} = b \times ln(10) \times y$$

This can be rewritten as:

$$\frac{dy}{y} \approx b \times ln(10) \times dx$$

If we consider a small change in x, i.e. $\delta x = 1$, then the approximate relative change in y is

$$\frac{\delta y}{y} \approx b \times ln(10) \times \delta x$$
$$\delta y \approx b \times ln(10) \times 1 \times y$$
$$\delta y \approx b \times ln(10) \times y$$

Hence, the percentage in y for a unit change in x is determined as $b \times ln(10) \times 100\%$. The sign of b determines the direction of change: negative for a decrease and positive for an increase.

A.4 Decision Tree Model

Figure 24 illustrates the top 3 levels of a Decision Tree Regressor trained to predict car prices.

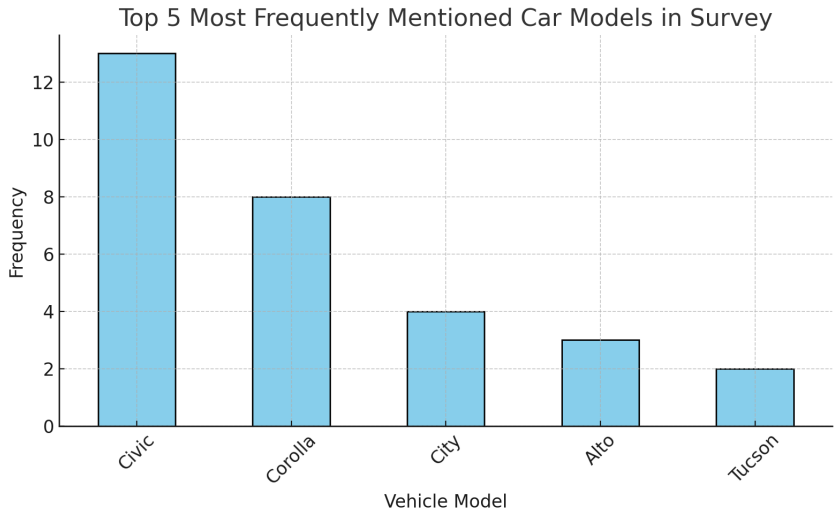


Figure 25: Most popular cars among respondents

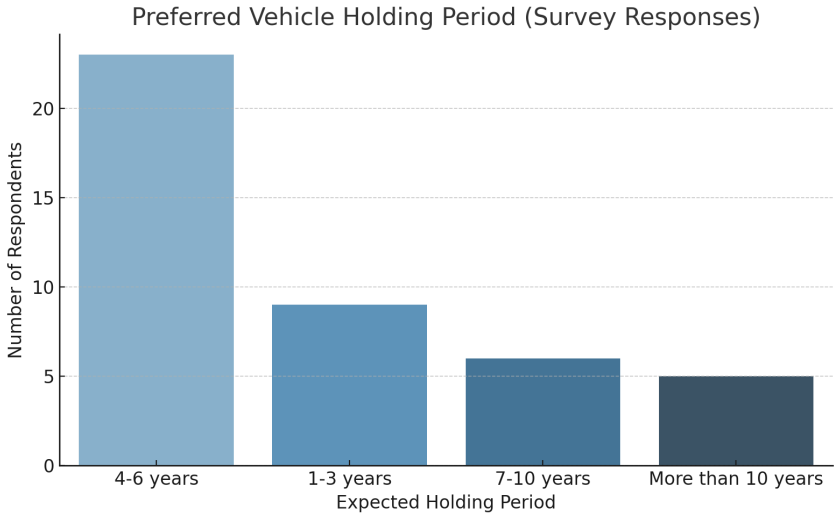


Figure 26: Expected holding period

A.5 Survey Insights

We conducted a survey to estimate the costs associated with owning a vehicle. As holding costs have already been discussed in the main body, here we will present some additional statistics from the survey. Figure 25 shows that the Honda Civic is the most popular car among respondents, followed by the Toyota Corolla.

Similarly, the majority of respondents prefer to keep their vehicle for 4–6 years, as shown in Figure 26, making it the most common holding period. Very few plan to hold a car for more than 10 years, indicating a trend towards more frequent resale.

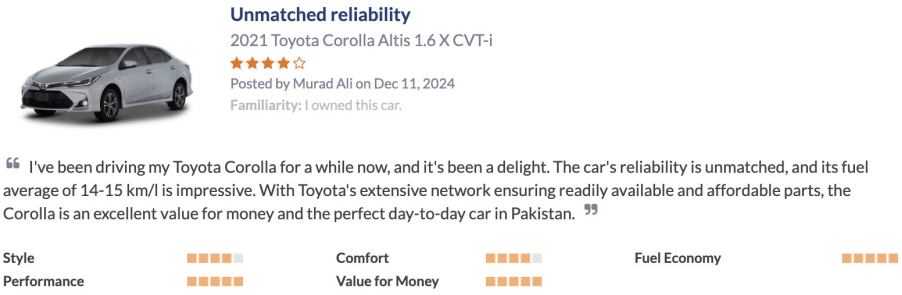


Figure 27: An example user review from PakWheels.com, highlighting key performance indicators: Style, Comfort, Fuel Economy, Performance, and Value for Money.

A.6 Optimization Model Details

A.6.1 Parameters

Here's an overview of the parameters used in the MOLP and the objective functions when calculating the optimal T_i, which were extracted from our dataset and whose origins have not been discussed previously.

- Average price of a make-model-year was calculated by taking the mean value of their current listings from the March 15th ad listings dataset.
- User Satisfaction and Value Efficiency scores were calculated by taking the average of Style, Comfort, and Performance Scores for the first, and the average of Fuel Economy and Value for Money ratings for the second.
- Popularity scores were calculated by taking the log-10 of the number of listings for each make-model. Since this was a proxy for resale potential, cars with the most listings should be easier to sell. A log transformation was used to account for outliers, such as the Toyota Corolla having listings orders of magnitude higher than, e.g., the Changan Alsvin.

A.6.2 Formulation

Finding T_i associated with each goal is simply a matter of setting the goal i as the objective function. With this subproblem, we are only interested in knowing the objective value, which we will use in our MOLP as a target and try to minimize undesirable deviations from it. Table 5 summarizes the results.

- The objective function would be the sum product of the parameters associated with the goal and decision variables that correspond to it, specifically z_k for the holding costs and y_j for the rest. Depending on the goal, it will be either MIN or MAX, as stated in Table 3.
- The constraints would be the same as those mentioned earlier in Section 5.2.6.

OBJECTIVE	TIER 1	TIER 2
T_1	95000	80000
T_2	4.8810	4.5581
T_3	4.7436	4.4326
T_4	0.5095	0.0188
T_5	3.8186	3.8186

Table 5: Optimal T_i associated with goal i for each budget tier in Section 5.2.7

A.6.3 Buyer Profiles

GOAL	PRIORITY LEVEL	BASE WEIGHT
User Satisfaction	1	2
Value Efficiency	1	1
Popularity	2	1
Depreciation	3	1
Holding Costs	3	1

Table 6: Premium-conscious buyer profile

GOAL	PRIORITY LEVEL	BASE WEIGHT
User Satisfaction	3	1
Value Efficiency	3	1
Popularity	1	2
Depreciation	1	1
Holding Costs	2	1

Table 7: Resale-oriented investor profile

GOAL	PRIORITY LEVEL	BASE WEIGHT
User Satisfaction	3	1
Value Efficiency	3	3
Popularity	1	1
Depreciation	2	1
Holding Costs	3	3

Table 8: Liquidity-seeking budget strategist profile

GOAL	PRIORITY LEVEL	BASE WEIGHT
User Satisfaction	1	1
Value Efficiency	1	2
Popularity	2	1
Depreciation	2	1
Holding Costs	3	1

Table 9: Value-oriented quality seeker profile

A.6.4 Step-by-Step MOLP Solution Procedure for Example Scenario from Presentation

This section gives a brief overview of how to use our model to get an optimal car recommendation.

Figure 28 shows a reddit post in which a user is looking to upgrade their car. We used an LLM[5] to assign priorities and weights for each of these goals based on their needs: build quality, fuel average, and comfort.

Prompt You are given a potential car buyer's profile describing their preferences. Based on this profile, your task is to assign:

1. Priority Levels (1 to 3): with 1 being the most important and 3 the least.

[5]GPT-4o was used for this task.

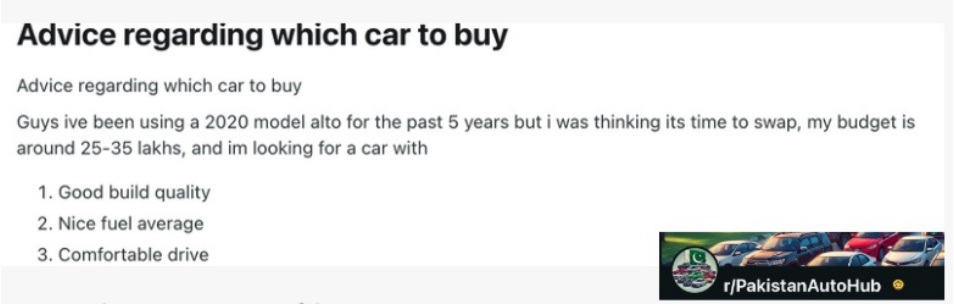


Figure 28: An example post from reddit community r/PakistanAutoHub.

2. Base Weights: to differentiate the relative importance of goals within the same priority level.

We are using a pre-emptive goal programming approach, so goals with higher priority levels must be satisfied before optimizing lower ones.

Refer to the following definitions for each goal:

1. Holding Costs: Maintenance, insurance, tax — costs incurred for keeping the car
2. User Satisfaction: Derived from review-based scores on comfort, reliability, features, etc.
3. Value Efficiency: Derived from ratio of satisfaction to total ownership cost, and fuel efficiency
4. Depreciation: Quarterly % decrease in price; lower means better long-term value retention
5. Popularity: Based on number of listings in the dataset; vehicle that is likely to be most easily sold

Here's an example for you to follow:

Buyer Profile:

```
<buyer_profile>
    Premium-Conscious Buyer: Such a buyer prioritizes comfort, luxury, prestige,
    and satisfaction over cost. They prefer stable long-term ownership and
    are less concerned about associated operational costs and depreciation.
</buyer_profile>
```

Priority Levels and Weights:

```
<weights>
    Goal-Priority Level-Weight
    Holding Costs-3-1
    User Satisfaction-1-2
    Value Efficiency-1-1
    Depreciation-3-1
    Popularity-2-1
</weights>
```

Now consider the buyer's profile:

```
<buyer_profile>
  Guys I've been using a 2020 model Alto for the past 5 years but I was thinking it's
   time to swap, my budget is around 25-35 lakhs, and I'm looking for a car with
    1. Good build quality
```

```
    2. Nice fuel average
    3. Comfortable drive
</buyer_profile>
```

Write your answer in the following format:

```
<holding_costs>
    Goal-Priority Level-Weight
</holding_costs>
```

Response Based on the prompt, the LLM assigns the following priorities and weights:

GOAL	PRIORITY LEVEL	BASE WEIGHT
User Satisfaction	1	1
Value Efficiency	2	5
Popularity	2	1
Depreciation	3	1
Holding Costs	3	1

Table 10: Reddit user buyer profile

Setting up the Excel Workbook The '*Model.xlsx*' file contains our formulated model[6], which can be used to get the optimal recommendation. The sheet named '*MOLP*[7]' is already populated with the parameters and constraints. The only thing we need to change is to edit the constraints labeled pink to match our problem.

constraints	LHS	Sign	RHS
Holding Costs	0	=	
User Satisfaction	0	=	
Value Efficiency	0	=	
Depriciation	0	=	
Popularity	0	=	
Budget Upper	0	<=	3500000
Budget Lower	0	>=	2500000
No Suzuki Alto	0	=	0
One Vehicle Selected - MMY	0	=	1
One Vehicle Selected - VT	0	=	1
Goal 4 Deviation	0	<=	0
Goal 5 Deviation	0	<=	0

Figure 29: Model constraints: green (goals), pink (problem-specific), orange (linking + one vehicle selection), and yellow (deviational variables).

- In this scenario, the buyer already owns a Suzuki Alto and is looking for an upgrade. Therefore, we will set the make-model decision variable corresponding to the Suzuki Alto to 0.
- The buyer's budget is between 25 lakhs and 35 lakhs. To represent this, two additional constraints will be added for the upper and lower limits. The upper limit budget constraint is already set; we just need to edit the RHS to 35 lakhs. Then, we

[6]This model is based on MINIMAX objective function.
[7]The solver requires this sheet to be the first sheet in the Excel workbook.

```
Option for printingOptions changed from normal to all
Total time (CPU seconds):       0.02   (Wallclock seconds):       0.03

Status: Optimal
Objective Value: 1500.06
Changan_Alsvin = 1.0
Changan_Alsvin_2019 = 1.0
Q_ = 1500.0
Sedan = 1.0
Y_2019 = 1.0
d_1p = 15000.0
d_3n = 0.37510804
d_4p = 3.5448248
d_5n = 1.5560727
```

Figure 30: The solver will output the minimax variable, the non-zero decision variables corresponding to vehicle type, make-model, year, make-model-year, and the non-zero deviational variables for each goal. Note that the objective value is meaningless to us.

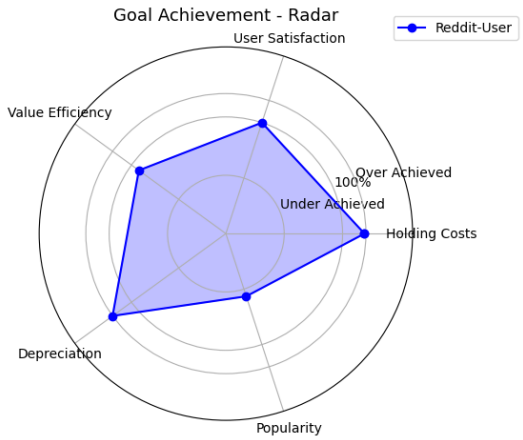


Figure 31: Goal achievement radar for example scenario

will copy and paste the same constraint, but this time change the sign to >= and the RHS to 25 lakhs.

- Next, add the priorities and base weights in the sheet named '*Penalties*'.
- Lastly, open the '*Solver.ipynb*' file and run the cell named '*Find Solution*'.

Constraints Color-Coding Figure 29 shows different color-coded constraints. Each of them is explained below:

- The green constraints represent the soft constraints or goals, and their RHS is intentionally left empty because we need to infer the optimal T_i via a separate LP with that goal as the objective function. Our solver will handle this.
- The pink constraints are problem-specific.
- The orange constraints represent the linking and one-vehicle selection constraints.
- The yellow constraints represent those for deviational variables.

Output For this buyer profile, our model suggests **Changan Alsvin 2019**. This choice aligns with their goals, having minimal deviations for the most prioritized goal as showcased by the goal achievement radar in Figure 31.