ENSF 612 - Engineering Large Scale Data Analytics Systems

Term Project Proposal

Project Title:

Second-Hand Vehicle Price Prediction Pipeline

Group Member	Name	UCID
Member 1	Mark Racca	30298613
Member 2	Ranjit Singh	30301717
Member 3	Edmund Yu	10124499

Submission Date:

November 7, 2025

Execution Platform:

Databricks (Free Tier)

Second-Hand Vehicle Price Prediction Pipeline

Engineering Problem

Develop a scalable ML pipeline to predict second-hand vehicle prices in the Toronto market, addressing real-world pricing accuracy challenges. The solution will handle diverse features, missing data, and process thousands of records efficiently.

Tools & Platform

Component	Technology	Purpose
Execution Platform	Databricks Community (Free)	Notebook environment, collaborative workspace
Data Processing	PySpark	Distributed DataFrame operations
Feature Encoding	StringIndexer/OneHotEncoder	Categorical feature transformation
Feature Scaling	StandardScaler	Numerical feature normalization
Storage & Pipeline	Delta Lake on Databricks	ACID transactions, medallion architecture

Machine Learning Models

Model	Approach	Purpose
Linear Regression	Simple linear relationships	Baseline performance
Random Forest	Ensemble of decision trees	Non-linear patterns
Gradient Boosting	Sequential tree boosting	Enhanced accuracy

Dataset Description

Source: Kaggle - Used Vehicles (Toronto 2023, Farhan Hossein) | **Records:** 24,199 vehicles | **Geography:** Toronto area (within 25km of downtown) | **Features:** Price (target), make, model, year, mileage, condition, fuel type, transmission, body type, drivetrain from Autotrader.ca

Big Data Engineering Relevance

Concept	Implementation
Distributed Processing	PySpark patterns on Databricks (scalable to larger data)
Pipeline Architecture	Medallion pattern (Bronze→Silver→Gold) on Delta Lake
Data Reliability	ACID transactions for versioning & governance
Scalability	ML pipeline extensible to larger datasets beyond free tier
Reproducibility	MLflow experiment tracking on Databricks