USED VEHICLE PRICE PREDICTION

Sanjay Jaras, Satish Agrawal

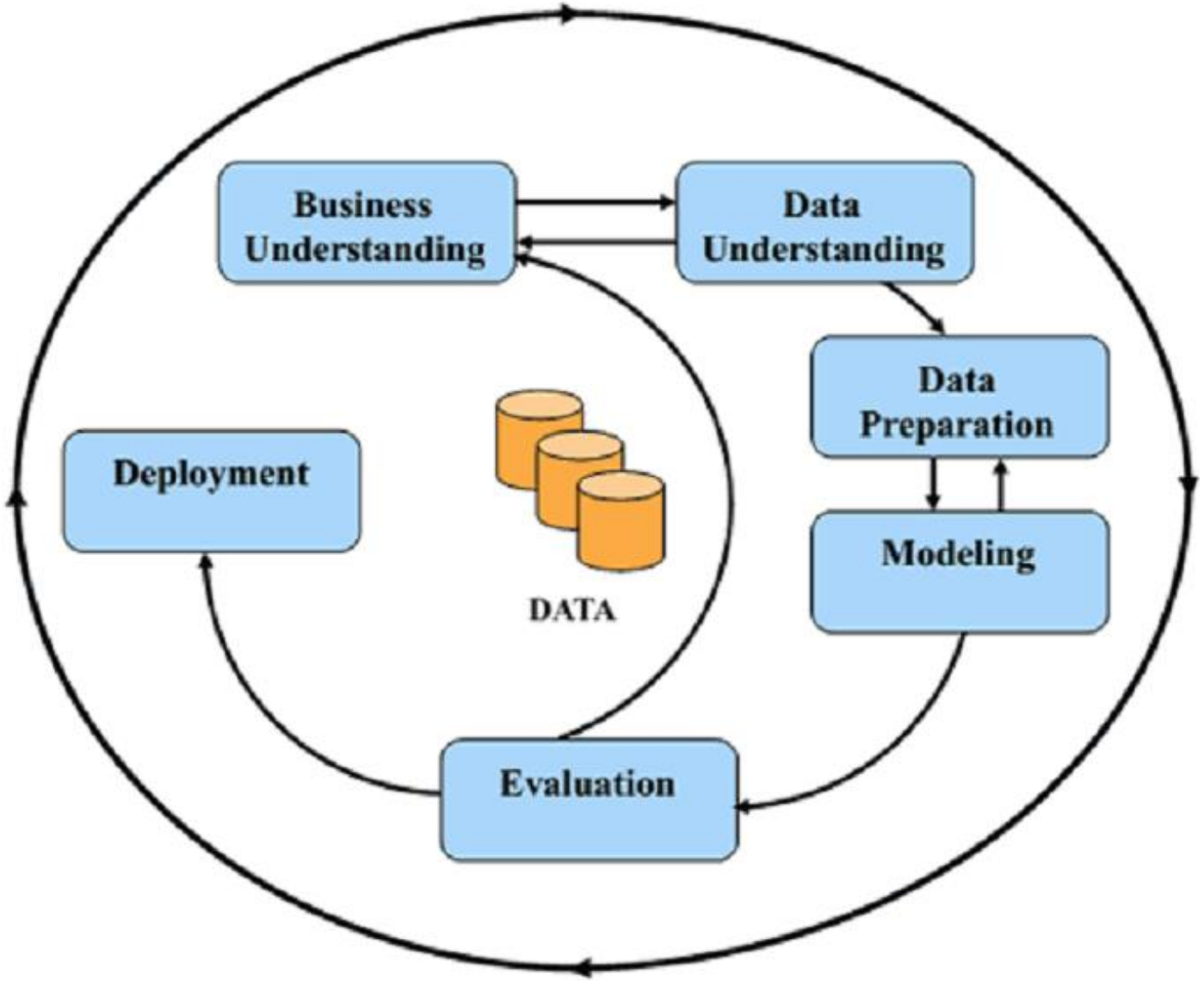
DSC-630

Bellevue University

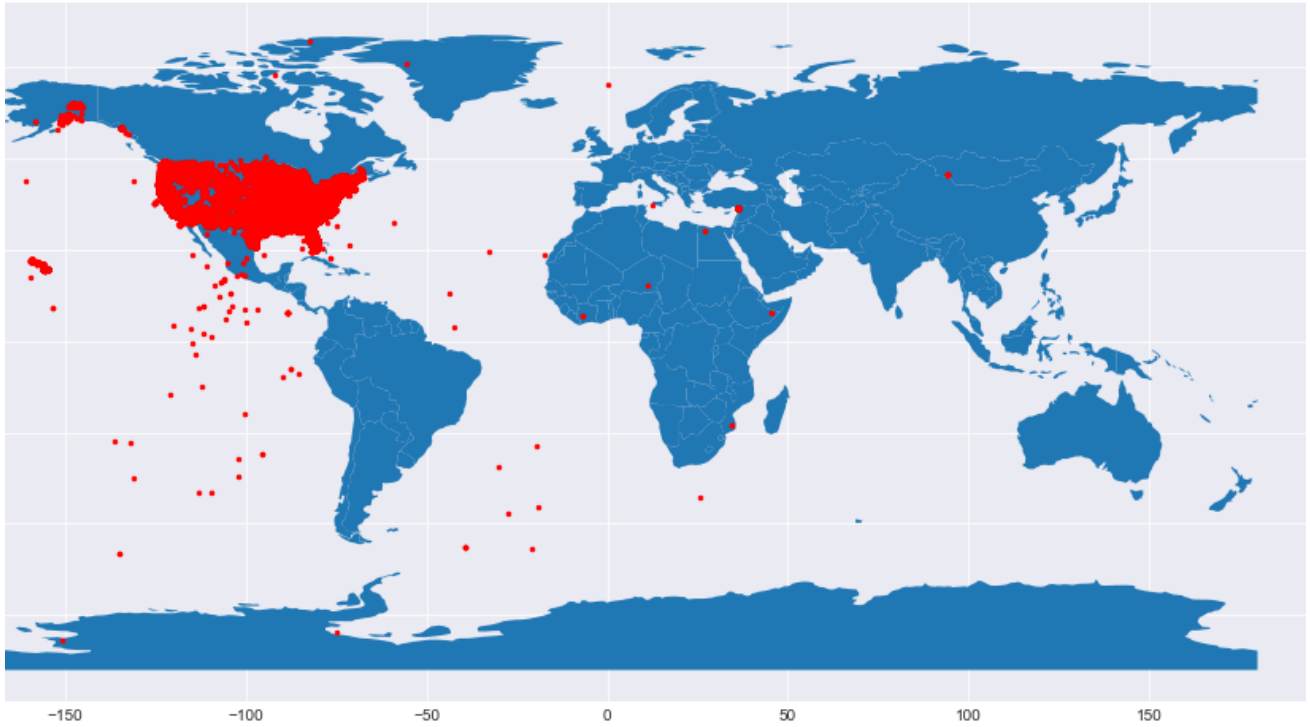
INTRODUCTION AND PROBLEM STATEMENT

- Price of a used vehicle is very important and needs to be consistent and fair.
- There are a variety of factors that impact the price of a used vehicle.
- This project uses a dataset with various features about the vehicles listed and sold in the past and attempts to predict a price for new listings.
- This will be useful in coming up with a fair price of the used vehicles, keeping the margin and market conditions in mind.

MAJOR STEPS

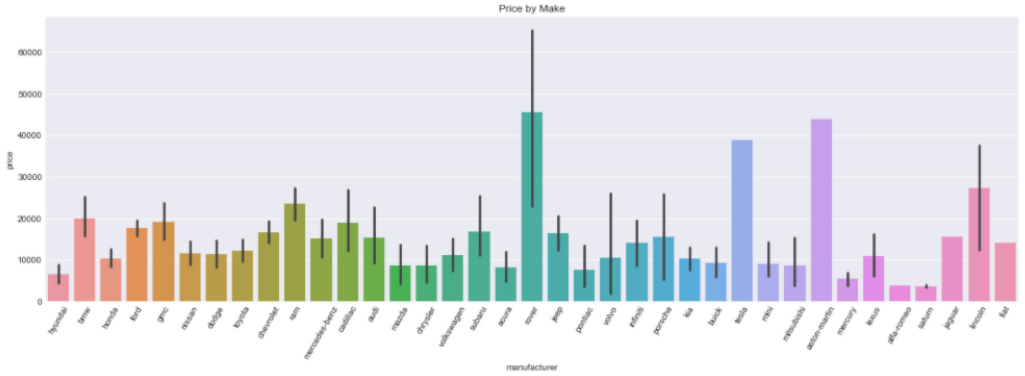


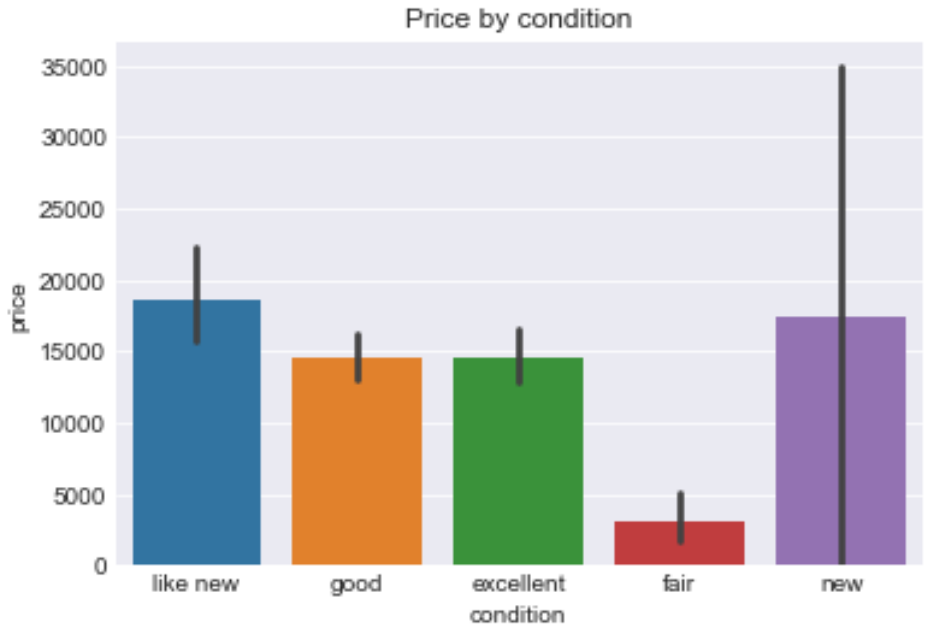
- Data understanding
- Split the data set into train and test set
- Exploratory analysis of the data set
- Data cleaning and preparation
- Feature selection and feature engineering
- Exhaustive model training
- Model evaluation and selection
- Testing with test data
- Discuss the best model and Model deployment



DATASET

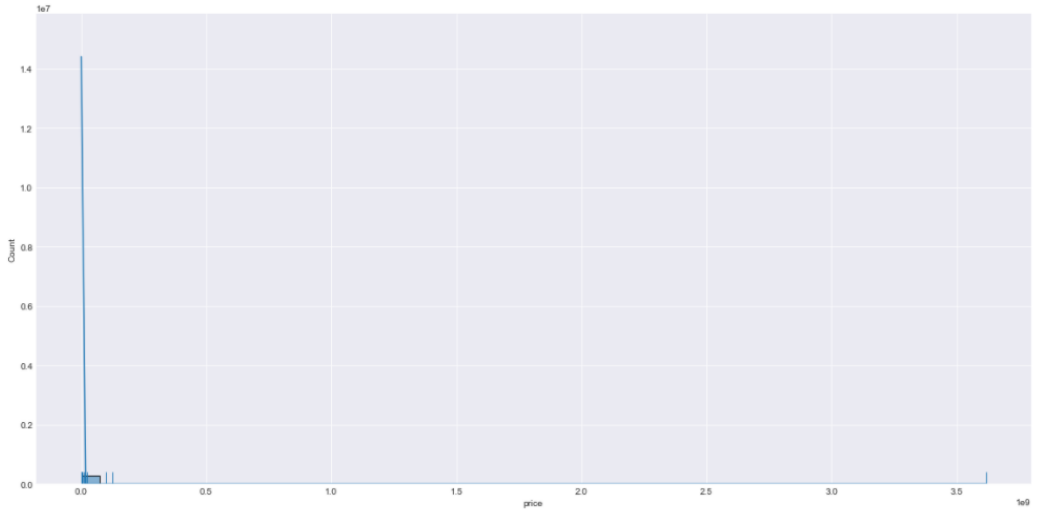
- The dataset used in this analysis is from Craigslist and has more than 458K observations of used vehicle listings.
- There are 25 attributes in the data set for each listing.
- The price for the vehicle is in the dataset and is our target variable that we are trying to predict using machine learning models.
- Key attributes are odometer, make and model, year, fuel type, and transmission.
- We kept 20% of the data aside for testing the model trained on 80% of the dataset.

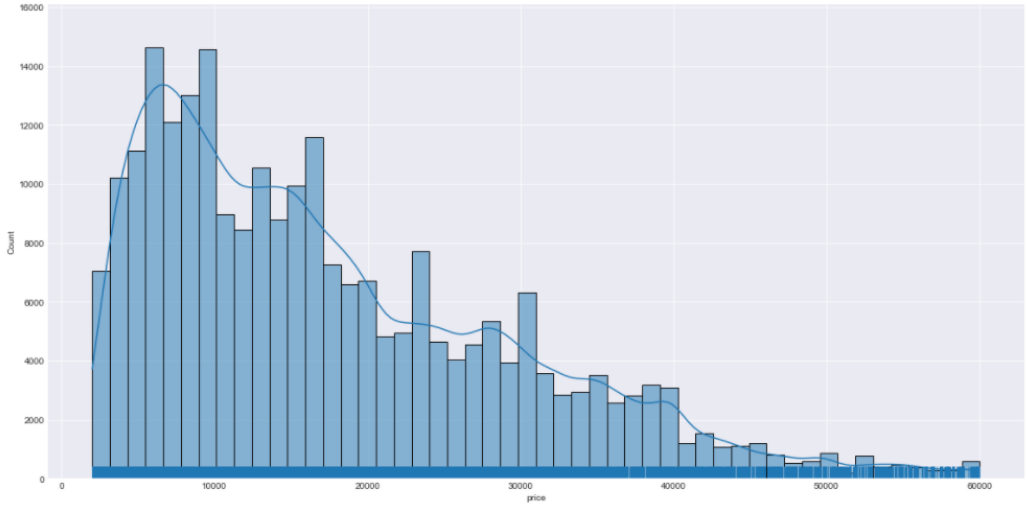




EXPLORATORY DATA ANALYSIS

- We performed EDA on most of the attributes to understand the distribution and find any correlations with the target variable.
- The two figures here show the correlation of condition and make with the target variable, the price of the vehicle.
- We generated a correlation matrix between all the numeric attributes to see the impact of individual variables on the target variable.
- We analyzed values for categorical variables for anomalies and inconsistencies.





DATA CLEANING AND PREPARATION

- We removed the outliers from price, year, and odometer.
- Above graphs show the distribution before and after outlier removal for price.
- We identified and imputed missing values using iterative imputer with ExtraTreesRegressor.
- We also removed the feature "size" because it has more than 50% missing values and may not be as helpful in price prediction.

FEATURE SELECTION AND ENGINEERING

- We removed the following attributes from the dataset, considering they do not have any correlation with the target variable or highly correlated with other features and are duplicates :
  - URL
  - ID
  - Image_URL
  - Description
  - Region_URL
  - VIN
- These features do not have any correlation with the price of the car and may not contribute towards making the price prediction
- We used model year and listing date to derive the age of the car at the time of listing. Age has an important role to play in price prediction.

TRAINING MODELS

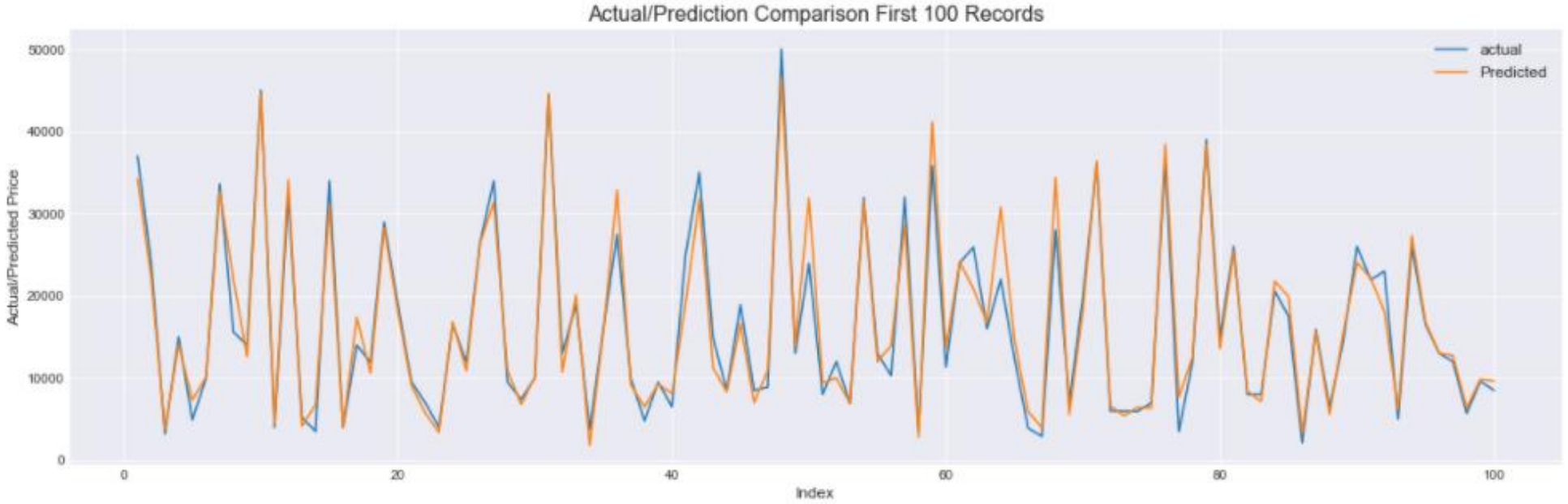
- Training is the process of applying available data to the chosen algorithm(s).
- We chose to train Linear regressor, DecisionTreeRegressor, XGBRegressor, and RandomForestRegressor models.
- Training multiple models allow comparing and choosing the best performing model.
- We employed an exhaustive search technique to train multiple models for the same algorithm and tune the hyperparameters to come up with the best model.
- We ended up training 85 models in total with cross-validation.

MODEL SELECTION AND SCORING METRICS

- Model selection is the process of selecting the best performing model with the best hyperparameters
- Scoring metrics allow comparing assorted models.
- There are many scoring metrics to choose from including Max error, r-square, explained variance, accuracy, F1 score, etc.
- We chose accuracy to be the scoring metrics for our model evaluation and selection process.
- K-fold cross-validation splits data into K-folds and uses K-1 folds to train a model and remaining one-fold to validate the performance using the metric provided.
- We used 5-fold cross-validation.

MAJOR CHALLENGES AND RESOLUTIONS

- Though the dataset has around 485k observations only, we have many categorical features with a long list of supported values, one hot encoding has increased the number of attributes in the set and was crashing the program while training.
- Dataset has many categorical attributes. Using one-hot encoding was making model overfit.
- As we split the dataset before in hand, we saw missing categories in the training dataset which were present in the test dataset.
- To avoid the above issue, we delayed the split of the dataset which was potentially introducing data snooping issues to the project.
- We found that sklearn has a newer version 0.24 where this has been handled and it worked fine for us.



OUTCOME AND NEXT STEPS

- At this point of the analysis, we have achieved 93% accuracy in used vehicle price prediction.
- Graph above shows predictions are accurate or close to the actual price for the vehicles.
- We are planning to work towards using PCA/SelectKBest for selecting the best correlated attributes to achieve better performance.
- We are exploring options to use Model ensembles to gain better accuracy.
- If we have other features like improvements to the vehicle or accessories added, would help improvise the prediction.