Project Title: Autozen Valuation Guru

Project Proposal

Team Members:

- Mengjun Chen
- Dhruvi Nishar
- HanChen Wang
- Tony Zoght

Partner:

Autozen Technology

Mentor:

Dr. Gittu George

Course:

DSCI 591 Capstone Project

Summer 2023

The University of British Columbia

Word count:

1199

Executive Summary

Autozen is a marketplace that connects private car sellers and dealerships in the used car market. They use the Canadian Black Book (CBB) to estimate used car values, but face pricing inconsistencies. Autozen suggests the Autozen Valuation Guru project, which uses Data Science and Machine Learning to understand how used car features affect prices, assess the current valuation method, and create a new valuation pipeline. The goal of this project is to improve used car valuation accuracy and precision. The project aims to create a reproducible pipeline and interactive dashboard. The project will use databases with 8 categories and 218 characteristics to help value used cars. This project can offer a better solution for buyers and sellers to make informed decisions with confidence, positioning Autozen as a leader in the used car market.

Introduction

Autozen is a marketplace that connects private car sellers and dealerships in the used car market ("Autozen," n.d.). Trading involves online valuation, inspection, auction, and offer. Autozen aims to simplify car selling and offer fair prices.

Autozen uses the Canadian Black Book (CBB) to predict used car market value ("Canadian Black Book," n.d.). Autozen struggles with pricing inconsistencies between estimated and final auctioned prices (Figure 1), hindering market confidence and growth potential. Autozen needs a Data Science solution to enhance used car valuation accuracy and reliability.

Autozen Valuation Guru will propose new Data Science (DS) and Machine Learning (ML) techniques to create a more accurate valuation pipeline for used cars. The project will analyze the relationship between used car features and auctioned price, evaluate the current valuation method, and create a new valuation process. The project aims to create a DS solution to estimate used car market value more confidently and reliably.

Autozen Valuation Guru is a crucial project that can aid the growth of the used car market by providing reliable valuations. Using DS and ML techniques, this project can provide a better solution for buyers and sellers, boosting Autozen's position as a used car market leader.

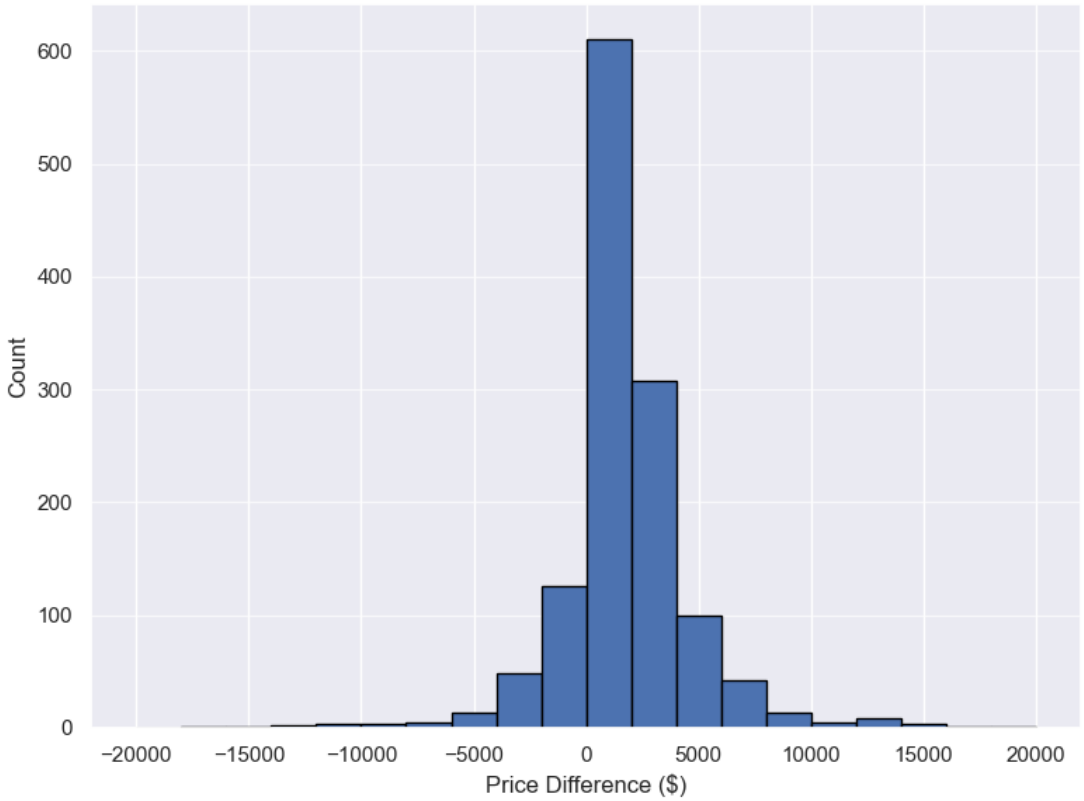

Figure 1: Differences between CBB predicted average and final auctioned price.

Project Description

Goals

This project uses DS and ML techniques to improve Autozen's used car valuations. The project seeks to:

1. Examine how used car features affect auction prices.
2. Assess how well the Canadian Black Book (CBB) car valuation method predicts the auctioned price.
3. Develop a predictive model for the final auctioned price using only the car inspection data without relying on CBB.

This project also has secondary goals:

1. Preprocess, clean, impute, and remove null values, normalize and scale numerical values, and convert Excel and JSON files to CSV.
2. Perform EDA to find data trends and patterns.
3. Assess Autozen's CBB-based valuation method.
4. Perform feature selection to identify the key features of a used car that explain the final auctioned price.
5. Perform feature engineering to include information such as the location, date, and season of the sale and investigate their contribution to the auctioned price.
6. Test new machine learning models for auction price prediction.
7. Create a reproducible project pipeline.
8. Use Tableau to create a data-visualization dashboard.

Stakeholders

The project has the following stakeholders:

Stakeholders	Interests	Impact on design
Autozen (primary)	Increase competitiveness in the market and user engagement.	Requires an accurate and precise prediction interval.
Sellers	Maximizing the sale price	Requires an accurate upper bound of prediction interval.
Buyers	Getting the best deal on a car.	Requires an accurate lower bound of prediction interval.

Data Science Techniques

We will start with basic methods and models to set a baseline, then increase complexity for better predictions. We will assess data suitability and potential obstacles for proposed methodologies below. The raw data originates from a read-only copy of the Autozen website, where sellers list and auction their pre-owned vehicles. The source data consists of three relational tables with a mixture of structured columns and JSON objects; we will process and extract the relevant data from these tables, which are vehicle inventory, vehicle inspections, and auction tables.

The tables contain information regarding car listings, inspections, and auctions, respectively. There are a total of 218 characteristics that aid in predicting its used-market value. These characteristics are divided into 8 categories:

- General information: make, model, year, mileage, color, and location, etc.
- Outside: rust status, scratches etc.
- Tires: tread depth, condition of rims, etc.
- Under-vehicle: exhaust state, differential etc.
- Under the hood: including the condition of the drive belt, battery, engine oil, etc.
- Lighting: working conditions of headlights and taillights, etc.
- Interior: seats, dashboard, odors, etc.
- Test drive evaluations: handling, braking system, external noise, etc.

The observational unit is the car listing and inspections. The auctioned price (whether accepted or not by the seller) represents the target to predict.

We will focus on data quality and standardization throughout the data preparation step to ensure reliable analysis. We will clean, impute, normalize, and scale numerical values and convert data (including the JSON objects) to CSV formats. These steps establish the groundwork for consistent and reliable analysis throughout the project.

We will select important features for training by analyzing the correlation between each feature and the auctioned price and using a number of techniques to reduce the dimensionality of the features. Such methods include **Group Lasso** which works well when there is a group of related features. Additionally, we will examine other PCA-based techniques like Multiple Correspondence Analysis (**MCA**) and Factor of Mixed Data (**FAMD**) which are both useful when there is a mixture of continuous, categorical, and ordinal features. We will also consider new features by combining and transforming existing ones to improve the prediction model's accuracy.

We will start with simple data science techniques to obtain a baseline for our prediction model, such as **linear regression**, Decision Trees and Ensemble Methods **(Random Forests, Gradient Boosting)**. We will evaluate the model's accuracy and compare it to the present Autozen (which is based on CBB) valuation method's accuracy in anticipating the auctioned price. As we improve the model, we will utilize more advanced techniques such as **neural networks**, and **deep learning**.

Difficulties and Challenges

The used car market is complex, and auction prices can fluctuate. Obtaining precise and dependable data can be difficult, and incomplete information can impact model accuracy. We will deal with these problems using techniques like data cleaning and imputation to ensure data accuracy and completeness. We must manage the partners' expectations as it's their first experience with Data Science and they have high expectations due to ChatGPT and LLMs hype.

Criteria for Success

Our goal is to improve the accuracy and precision of predicting auction prices. We will use MAPE and adjusted R-squared to assess model accuracy. Autozen aims to improve the prediction interval accuracy compared to the current CBB-based model.

Deliverable

1. A reproducible **Jupyter Notebook** containing the code and documentation.
2. A **Tableau dashboard** displaying essential features extracted from the EDA.
3. **Python scripts** for implementing our ML models with a Pipeline.
4. **Docker** for the deployment environment.

Project Timeline

Milestone 1 (Thursday, May 11th): Identify baseline model and evaluation metrics.

Milestone 2 (Thursday, May 18th): Assess initial models.

Milestone 3 (Thursday, May 25th): Evaluate the final model.

Milestone 4 (Thursday, June 1st): Assemble all deliverable.

Milestone 5 (Thursday, June 8th): Submit the final deliverable.

References

"Autozen." n.d. https://www.autozen.com/.
"Canadian Black Book." n.d. https://www.canadianblackbook.com/.