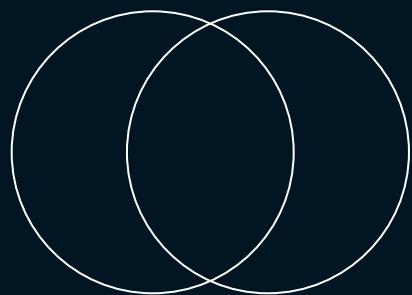


# Q1 - Week 1

# Group 2 Presentation



Ji-Soo Kim, Tae-Yoon Kim, Jun-Beom Lee, Jin-Joo Yang, Sung-Hyun Kim



# Table of Contents

<b>1. Topic Introduction</b>	_____
<b>2. Background of our topic</b>	_____
<b>3. Selected Dataset</b>	_____
<b>4. Preceding papers analysis</b>	_____
<b>5. Expected Effect</b>	_____
<b>6. Q &amp; A</b>	_____

## 1. Introduction

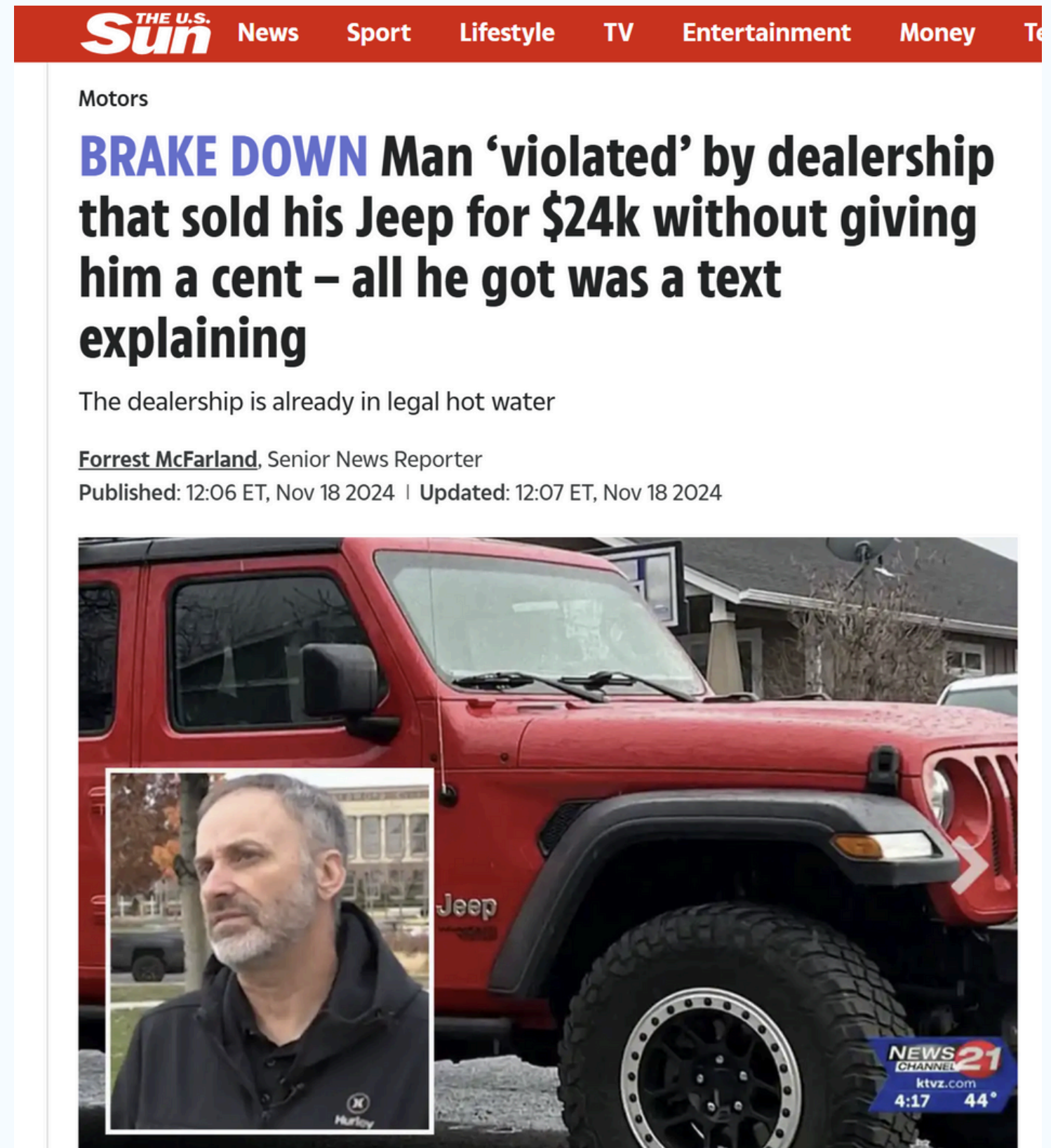
# Price Prediction of Private Sales Vehicles with Used Car Datas

## 2-1. Background of our topic

---

Jeff Peck, who commissioned a car sale to a dealer in Prineville, Oregon, in November 2024, **suffered the damage of not receiving his payment** after five months, even though his Jeep sold for \$24,000. The investigation found that the dealer not only failed to pay for the sale, but also committed **illegal acts**, such as forging documents and forging signatures.

---



## 2-2. Background of our topic



### “ Normal Index”

: the only real-time barometer of the used car market's recovery

**actual used car prices  
> expected prices**

☒ Copilot return to normal index  
( 2020-2022)

# 3. Selected Dataset

≡

kaggle

+

Create

🏠

Home

🏆

Competitions

📁

Datasets

🔗

Models

⏏

Code

💬

Discussions

🎓

Learn

⌵

More

📁

Your Work

▼

VIEWED

🔍

Search

AUSTIN REESE · UPDATED 4 YEARS AGO

Used Cars Dataset

Vehicles listings from Craigslist

Data Card

Code (215)

Discussion (20)

Suggestions (0)

About Dataset

Context

Craigslist is the world's largest collection of used vehicles for sale, yet it's very difficult to place. I built a scraper for a school project and expanded upon it later to create this dataset of vehicle entry within the United States on Craigslist.



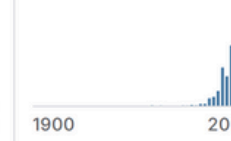
vehicles.csv (1.45 GB)

Detail

Compact

Column

26 of 26 columns

id entry ID	url listing URL	region craigslist region	region_url region URL	price entry price	year entry year	manufacturer manufacturer of vehicle	model model of vehicle
 7.21b7.32b	426880 unique values	404 unique values	413 unique values	 03.74b	 19002022	ford17% chevrolet13% Other (300831)70%	f-1502% [null]1% Other (413595)97%
7314910156	https://auburn.craigslist.org/ctd/d/auburn-university-2018-ford-f150-super/7314910156.html	auburn	https://auburn.craigslist.org	34590	2018	ford	f150 super cab x1 pickup 4d
7314854462	https://auburn.craigslist.org/ctd/d/auburn-university-2016-toyota-tacoma/7314854462.html	auburn	https://auburn.craigslist.org	30590	2016	toyota	tacoma double cab sr5
7314811916	https://auburn.craigslist.org/ctd/d/auburn-university-2020-jeep-wrangler/7314811916.html	auburn	https://auburn.craigslist.org	32990	2020	jeep	wrangler sport suv 2d
7314811909	https://auburn.craigslist.org/ctd/d/auburn-university-2020-ford-f150/7314811909.html	auburn	https://auburn.craigslist.org	38990	2020	ford	f150 supercrew cab xlt



## **3-1. Selected Dataset - column**

- 1. price : car enter price (USD)**
- 2. year : car entry year**
- 3. manufacturer : manufacter of vehicle (ex: Toyota etc.)**
- 4. model : specific model name**
- 5. condition : condition of vehicle (ex: new, good, etc.)**
- 6. cylinders : number of cylinders (ex: 6 cylinders etc.)**
- 7. fuel : fuel type (ex: gas, disel, electric etc.)**
- 8. odometer : miles traveled by vehicle**
- 9. title\_status : title status of vehicle (ex: clean, accidents etc.)**
- 10. transmission : transmission of vehicle (ex: automatic, manual, etc.)**
- 11. drive : type of drive (ex: 4wd, fwd, rwd)**
- 12. size : size of vehicle (ex: full-size)**

## **3-1. Selected Dataset - column**

**13. type : generic type of vehicle (ex: sedan, SUV, truck)**

**14. paint\_color : color of vehicle**

**15. lat : latitude of listing**

**16. long : longitude of listing**

**17. posting\_date : Date the vehicle information was published**

**18. url : listing url**

**19. region : craigslist region**

**20. region\_url : region URL**

**21. VIN : vehicle identification number**

**22. image\_url**

**23. description : listed description of vehicle**

**24. state : state of listing**



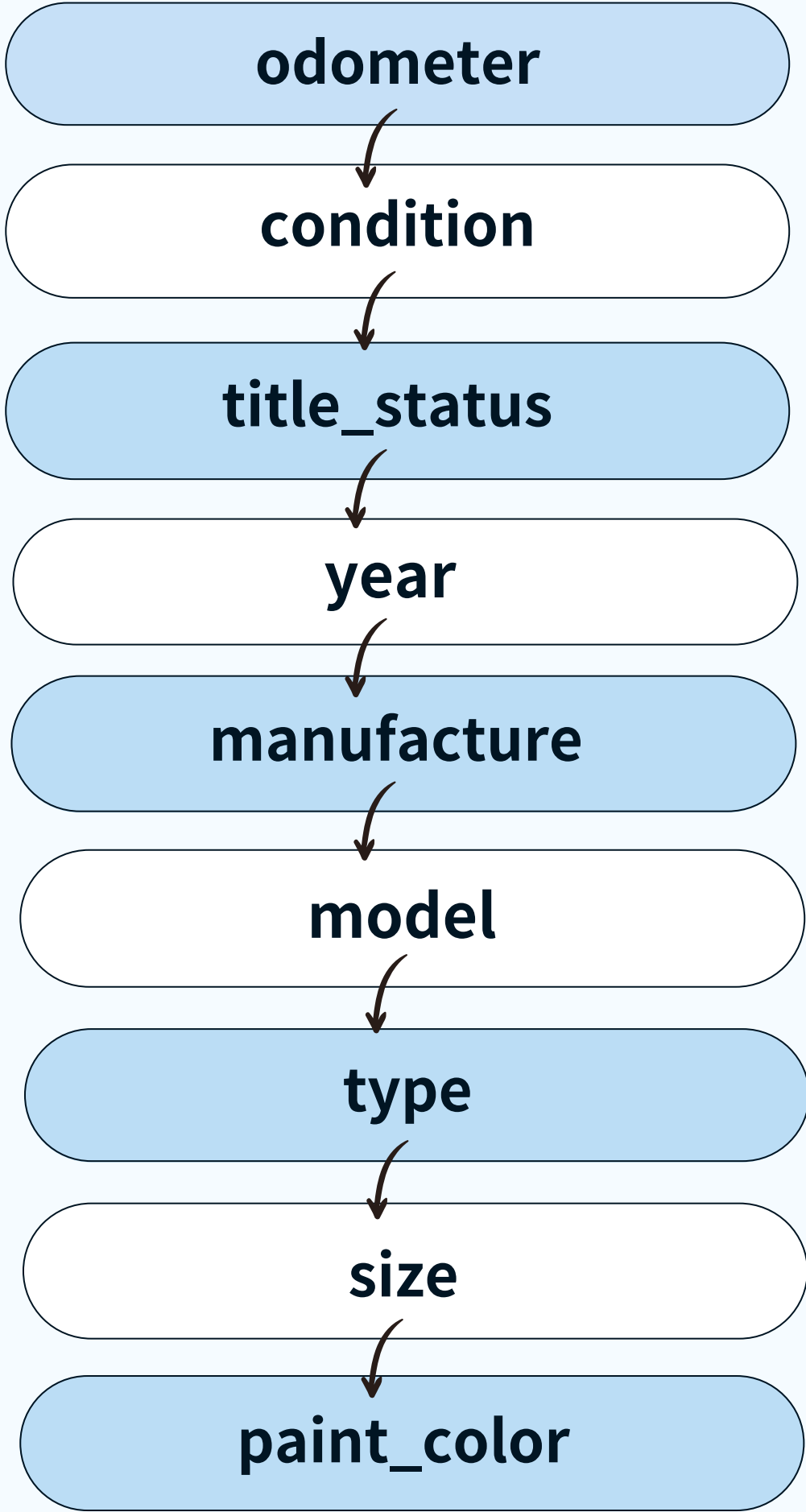
# 3-2. Selected column

	column		column		column		column
1	id	5	cylinders	9	drive	13	model
2	price	6	fuel	10	size	14	transmission
3	year	7	odometer	11	type	15	posting_date
4	manufacturer	8	title_status	12	paint_color	16	condition

### 3-3. Selected Columns ranking

Table 1. Ranking the Factors that Influence Customers Purchase of Imported used Cars

Factors	Mean	Standard Deviation	Rank
Car interior	3.52	1.98	8 <sup>th</sup>
Price	4.04	0.73	5 <sup>th</sup>
Mileage	4.52	1.00	1 <sup>st</sup>
Chassis condition	4.21	1.83	3 <sup>rd</sup>
Country of Origin	4.37	1.11	2 <sup>nd</sup>
Brand name	4.09	1.09	4 <sup>th</sup>
Dealers' reputation	3.12	1.88	9 <sup>th</sup>
Tires	2.82	1.99	11 <sup>th</sup>
Car documents	3.89	0.89	7 <sup>th</sup>
Car exterior	3.92	1.03	6 <sup>th</sup>
Colour	2.98	0.99	10 <sup>th</sup>



# 4. Preceding papers analysis

2023 International Conference on Electrical, Computer and Energy Technologies (ICEET 2023)  
16-17 November 2023, Cape Town-South Africa

### Predictive Analytics on Used Car Prices Using Business Intelligence of Bayesian Networks for Sales Risk Reduction

Jewel Donkor Apeko  
Department of Computer Science  
Norfolk State University  
Norfolk, VA, USA  
j.d.apeko@spartans.nsu.edu

Isaac O. Osunmakinde  
Department of Computer Science  
Norfolk State University  
Norfolk, VA, USA  
ioosunmakinde@nsu.edu

Musbah M. Abdulgader  
Department of Computer Science  
Norfolk State University  
Norfolk, VA, USA  
mmabdulgader@nsu.edu

Kingsley C. Nwosu  
Department of Computer Science  
Norfolk State University  
Norfolk, VA, USA  
kcnwosu@nsu.edu

**Abstract**— Used car prices are normally estimated at the time of initial purchase when the car is freshly purchased. Due to the shortage of cars from car manufacturers during the pandemic periods, secondhand car dealers have gained an undue advantage by withholding some information about cars advertised online. Researchers have tried to develop models to address this issue. There has not been enough research done in the area of uncertainty (partial observability) about used car prices. This research develops three transparent probabilistic models: a generalized model for all used cars, a specialized model for luxury cars, and a specialized model for non-luxury cars. It is intended to determine which models best predict used car prices. Publicly available car datasets were sourced, cleaned, and processed to build the models. Experimental evaluations reveal that the generalized model predicts prices with an accuracy of 80.19%, the specialized BN model for non-luxury cars predicts prices with an accuracy of 80.13% while the specialized BN model for luxury cars predicts prices with an accuracy of 83.54%. The results suggest that a specialized model performs better when predicting luxury used car prices. However, with non-luxury used cars, any of the two models predict prices satisfactorily. This research is aimed at helping used car buyers get a fair idea of expected selling prices, reducing risks using lower and upper bound prices. A new car buyer can also apply this model to estimate how much their car will be worth after a few years, which could aid in reaching sound business decisions.

**Keywords**— *Automobile, Bayesian Networks, Business Intelligence, Prediction, Probabilistic Modelling, Sales Risk Reduction, Used Car Price.*

#### I. INTRODUCTION

The price of a used car is normally estimated at the time of initial purchase when the car is new, using the residual value of the car which is based on the Manufacturer Suggested Retail Price (MSRP) of the car. However, there are so many factors that can alter the estimated used car price. In this research, a model is proposed to predict the prices of used cars while accounting for the uncertainties that arise from the unavailability or loss of data about a vehicle, which can significantly alter the estimated price of a used car.

The prices of used cars have been skyrocketing in recent times due to slow production of vehicles because of restrictions due to the COVID-19 pandemic [1] as well as a shortage in semiconductor chips [2] necessary for implementing modern functionalities such as automatic emergency braking, adaptive cruise control, lane keeping assistance among other features into modern cars. For instance, the shortage of computer chips has forced automakers to focus on producing only car models with high-profit margins. Available data indicates that the used car market contributes more to the overall car trade in the United States compared to new car sales. For instance, in 2020, 39.3 million used light vehicles were sold in the US while 14 million new light vehicles were sold [3]. This trend has given an undue advantage to used car dealers to unnecessarily increase used car prices and, in some cases, withhold essential information about cars they showcase online especially when this information could affect the price a potential customer would be willing to pay for the car. This research therefore intends to develop probabilistic models to efficiently predict the prices of used cars with complete or incomplete data.

#### A. Recent Customer Dissatisfaction Due to Poor Sales Decisions

Most car dealerships have typically put their profits ahead of customer satisfaction as evidenced by the increase in customer dissatisfaction during the COVID-19 pandemic (2020) when customers were forced to purchase cars online. Fig. 1 shows the percentage of car buyer satisfaction from 2016 to 2020 as reported by Statista [6].

This gives more credence to a 2016 survey conducted by Harris Polls that revealed that 61% of Americans feel they are taken advantage of when shopping at a car dealership and 81% dislike something about the car purchase process at car dealerships. The survey also revealed that 42% of Americans would be comfortable purchasing a car online without a test drive if certain assurances were in place. A key finding from JD Power revealed that online buyers are more satisfied with the car purchase experience than those who physically go to car dealerships [7].

979-8-3503-2781-6/23/\$31.00 ©2023 IEEE

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 16,2025 at 19:38:24 UTC from IEEE Xplore. Restrictions apply.

## Problems

During the pandemic, the used car market has grown rapidly due to production shortages from automakers, and uncertainty in pricing has increased as dealers unfairly manipulate or hide information.

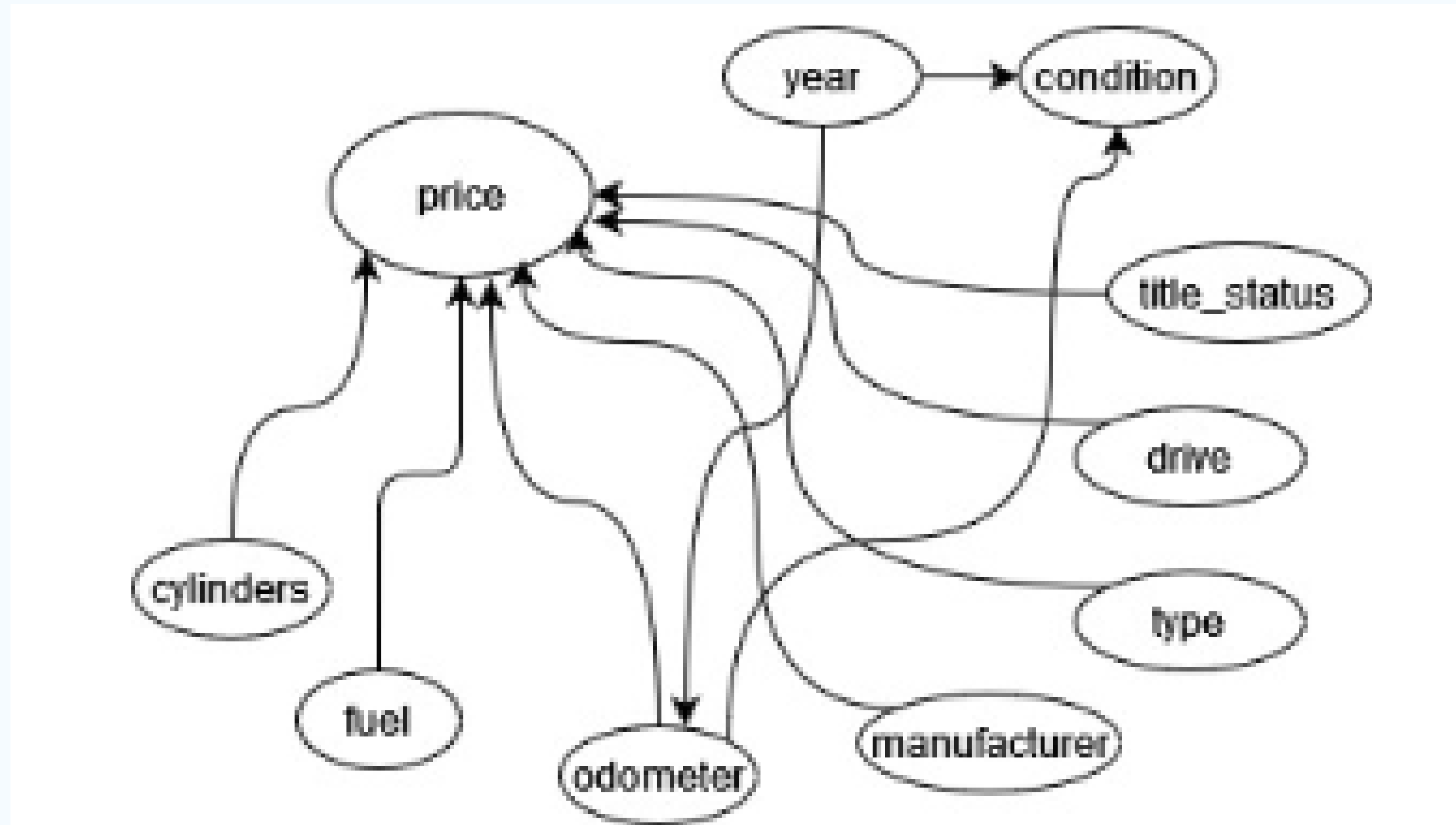
## Direction of this Paper

The accuracy of price prediction was compared and analyzed by developing a model specialized for each of luxury and general used cars.

11

## 4. Preceding papers analysis

---



A picture of a used data column from this paper



## 4. Preceding papers analysis

### C. Evaluation Metrics

The completed Bayesian model was evaluated by performing K-Fold Cross-validation and calculating the percentage error (using the Mean Absolute Error) and subsequently the accuracy. By definition, the Mean Absolute Error is the average of the absolute difference between the actual target value and the predicted target value. The evaluation metrics shown in equations (2), (3), and (4) were adapted from [15].

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (2)$$

$$\text{Percentage Error} = \frac{(\text{MAE} \times 100)}{\text{Average of Target variable}} \quad (3)$$

$$\text{Accuracy} = 100\% - \text{Percentage Error} \quad (4)$$

### G. Performance Evaluations

Table I presents the performance evaluations of the proposed new models using equations (2) – (4).

TABLE IV. PERFORMANCE EVALUATIONS OF THE BAYESIAN MODELS

<i>Model</i>	<i>Accuracy</i>
Generalized BN Model	80.19%
Specialized BN model (luxury vehicles)	83.5%
Specialized BN model (non-luxury vehicles)	80.13%

Due to the higher accuracy obtained on the specialized BN for luxury cars, this research suggests its deployment in real life than the generalized models.

## 4. Preceding papers analysis

---

### Significance

- Luxury vehicles require a separate predictive model because they have **larger initial price fluctuations** than regular vehicles and **target specialized demand groups!**

So based on this, we might think of....

### 1. Adding price range criteria

- Distinguishing between luxury and regular vehicles based on vehicle prices
- (e.g. \$50,000 or more is considered luxury)



## 4. Preceding papers analysis

---

### 2. Use other supervised learning algorithms

- With KNIME -> **75000 data samples** after preprocessing  
=> Since it's a medium~large dataset more diverse supervised learning algorithms are available, and there might be a good chance of **higher performance than Bayesian networks** used in this paper!

**Random Forest**

**SVR**

**Linear Regression**

**XGBoost**

### 3. Change the target!

- This paper's target : **CONSUMER** of the used car
- Our target : **OWNER** of the used car



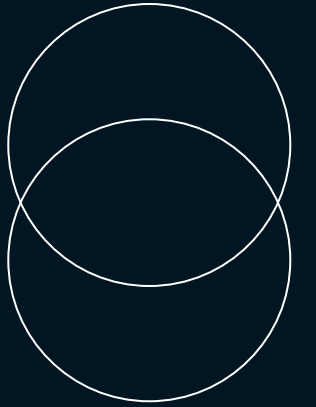
## 5. Expected Effect

**Support for Individual Sellers**

**Improved Market  
Transparency and Trust**

**Growth of the Used Car Market**

**Increased Transaction Efficiency**



# Q & A

Thank you for listening!

