Used Cars Purchase Risk Analyzation

Final Report for Advanced Data Resource Management 2023 Fall

Bi-Lian Chen †

Backstory and Purpose

Purchasing used cars at auto auctions poses a significant challenge for dealerships, mainly due to the inherent risk that a vehicle might harbor substantial issues, rendering it unsuitable for sale to customers. Within the automotive community, these unfortunate acquisitions are commonly referred to as "kicks."

Instances of kicked cars often arise from various factors, including odometer tampering, unresolved mechanical issues beyond the dealer's capability, complications in obtaining the vehicle title from the seller, or other unforeseen complications. For dealers, kick cars can lead to substantial costs, encompassing expenses for transportation, extensive repair work, and market losses during the resale process.

Modelers equipped with the ability to identify cars with a higher likelihood of being kicked can offer invaluable assistance to dealerships striving to curate the best possible inventory for their customers. The primary objective of this project is to find out what features can be main factors to predict whether a used car purchased is deemed a "kick" (a bad buy) or a good buy.

Executive Summary

Through this project, I use SQL to investigate into the risk factors associated with purchasing used cars to analyze whether a vehicle is likely to be a 'kick' or a bad buy based on their other features. I focused on variables such as odometer readings, cost, make, color, and age, the research has dissected over 70,000 vehicle transactions to uncover trends that could significantly impact dealership inventory decisions. After that, I learned how to use Excel to create Pivot Table and generate graphs to find the answers for my hypothesis and support my conclusions.

Hypothesis Before Start the Project

1. The higher the odometer reading the used cars have, the higher chance that they are badbuys.
2. The cheaper costs the used cars have, they are tend to be badbuys.
3. Vehicles color should not have big effect in the badbuy risks.
4. There should be certain makes that are more risker than other brands.
5. Vehicles age may have significant relationships with the used car risk.

6. Online shopping for used cars should have more risk than in person shopping.

Major Findings

1. Higher odometer readings correlate with an increased likelihood of cars being bad buys, particularly in the 40,000-60,000 and 60,000-80,000 miles ranges. The only exception are the cars in the range of 0-20,000 miles, the incidence of bad buys is unexpectedly higher than in the 60,000-80,000 miles category. This may suggest that vehicles with unusually low mileage appearing in the used car market could be indicative of serious underlying issues, such as accidents or malfunctions.
2. The cheapest cars ($0 - $2K) are most likely to be increases, with a clear trend showing decreased bad buy rates as vehicle costs increases.
3. Vehicle color does impact bad buy rates with purple vehicles having the highest bad buy rate at 15%.
4. Certain makes, notably DODGE, FORD, and MITSUBISHI, have higher instances of bad buys compared to others like HONDA and LEXUS, which have very low bad buy numbers.
5. Vehicle age is a significant factor, with the probability of a car being a bad buy increasing sharply after the 4-year mark.
6. Online purchased cars have a surprisingly low bad buy rate of 2.3%, challenging the hypothesis that online purchases are riskier.

Main Conclusions

1. The hypothesis that higher odometer readings and lower costs are indicators of potential bad buys is confirmed.
2. Contrary to initial assumptions, vehicle color does affect bad buy risks, with purple vehicles being more likely to be bad buys.
3. The hypothesis that certain makes are riskier is supported, although the data may be biased towards U.S. brands.
4. The assertion that vehicle age has a significant relationship with the risk of being a bad buy is validated, with older vehicles being riskier.
5. The hypothesis that online shopping for used cars carries more risk is rejected based on the data, indicating that online platforms may provide sufficient quality checks and transparency.

Dataset

Data is from Kaggle Competition called ‘Don't Get Kicked!’. It includes different features in over 70,000 used car purchases and the result that if each of them is referred to as a bad buy. Link https://www.kaggle.com/competitions/DontGetKicked. The

following column is a brief data dictionary on the dataset.

Field Name	Definition
RefID	Unique (sequential) number assigned to vehicles
IsBadBuy	Identifies if the kicked vehicle was an avoidable purchase
PurchDate	The Date the vehicle was Purchased at Auction
Auction	Auction provider at which the vehicle was purchased
VehYear	The manufacturer's year of the vehicle
VehicleAge	The Years elapsed since the manufacturer's year
Make	Vehicle Manufacturer
Model	Vehicle Model
SubModel	Vehicle Submodel
Color	Vehicle Color
Transmission	Vehicles transmission type (Automatic, Manual)
WheelTypeID	The type id of the vehicle wheel
WheelType	The vehicle wheel type description (Alloy, Covers)
VehMeter	The vehicles odometer reading
Nationality	The Manufacturer's country
Size	The size category of the vehicle (Compact, SUV, etc.)
TopThreeAmericanName	Identifies if the manufacturer is one of the top three American manufacturers
AcquisitionAuctionAverage Price	Acquisition price for this vehicle in average condition at time of purchase
AcquisitionRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition at time of purchase
CurrentAuctionAveragePrice	Acquisition price for this vehicle in average condition as of current day
CurrentRetailAveragePrice	Acquisition price for this vehicle in the retail market in average condition as of current day
PRIMEUNIT	Identifies if the vehicle would have a higher demand than a standard purchase
AUCGUART	The level guarntee provided by auction for the vehicle (Green light - Guaranteed/arbitratable, Yellow Light - caution/issue, red light - sold as is)
BuyerNumber	Unique number assigned to the buyer that purchased the vehicle

VehPurchState	State where the car was purchased
VehBCost	Acquisition cost paid for the vehicle at time of purchase
IsOnlineSale	Identifies if the vehicle was originally purchased online
WarrantyCost	Warranty price (term=36month and millage=36K)

Data Quality Check

In this project, I have checked if they have any irrational numbers in all 27 features. Most of them are collected correctly, only two features 'PRIMEUNIT' and 'AUCGUART' need to be cleaned before we can do any analysis.

'PRIMEUNIT' and 'AUCGUART' are having the same issue that they only have small amounts of valid data points. For example, 'AUCGUART' has only 3312 valid data out of 70612 total data points, 'PRIMEUNIT' also has similar numbers of valid data. Therefore, I transform their 'NULL' data points into '0' in the Fact table, which allows me to filter these out when I want to analyze these two features.

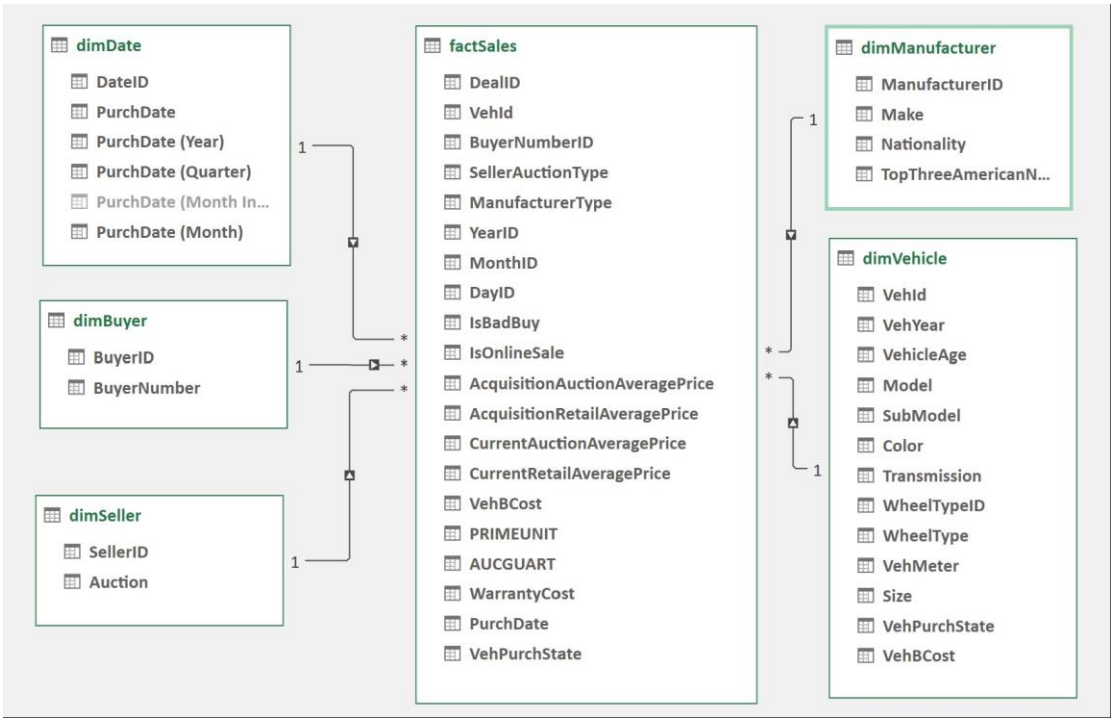
Data Organization

In the process of organizing the sales data within the database, I use Microsoft SQL Server Management Studio to construct the following tables:

- factSales: The central fact table records individual sales transactions, capturing details such as the sale's unique identifier (DealID), the vehicle involved (VehId), the buyer and seller information, and transaction specifics including the sale price and whether the vehicle was a bad buy (IsBadBuy).
- dimDate: This dimension table holds temporal details of each purchase, including the purchase date and its broken-down components like year, quarter, month, and day.
- dimBuyer: This dimension table contains information about the buyers, including a unique buyer ID and buyer number.
- dimSeller: This table provides data on the sellers, storing each seller's unique ID and the auction through which the sale was made.
- dimManufacturer: This dimension table catalogs information about the vehicle manufacturers, such as manufacturer ID, make, nationality, and if they are one of the top three American manufacturers.
- dimVehicle: The vehicle dimension table details the vehicles themselves, including IDs, year, age, model, submodel, and other relevant attributes such as color, transmission type, and size.

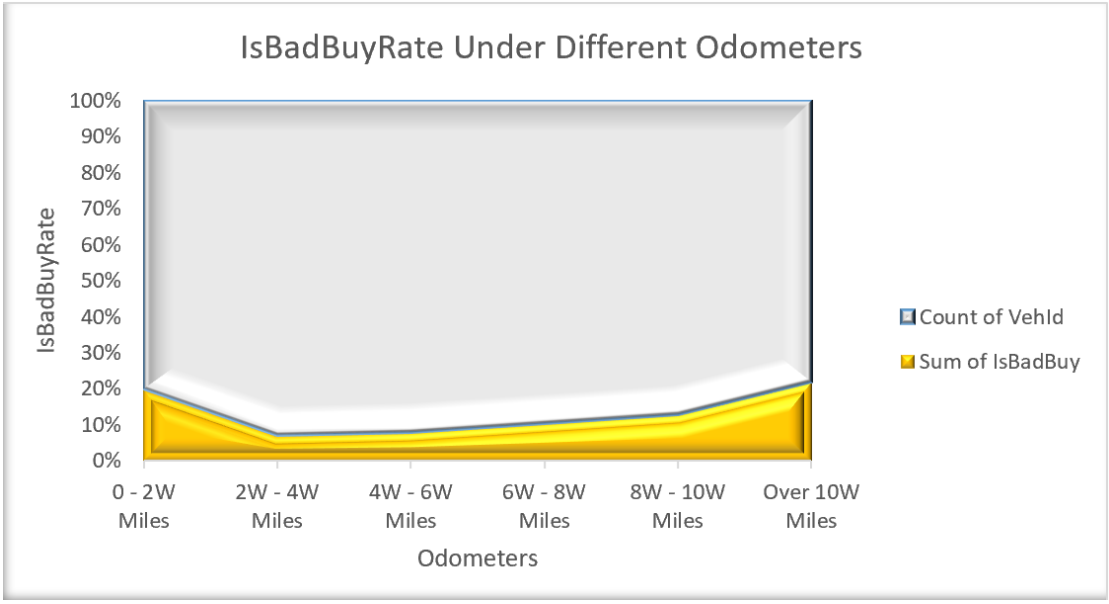
Each dimension table is linked to the central factSales table through foreign keys,

enabling comprehensive analysis across multiple dimensions such as time, buyer/seller profiles, vehicle details, and manufacturer information. This relational database structure is designed to facilitate complex queries for business intelligence and reporting purposes.



Visualization

1. IsBadBuyRate Under Different Odometers



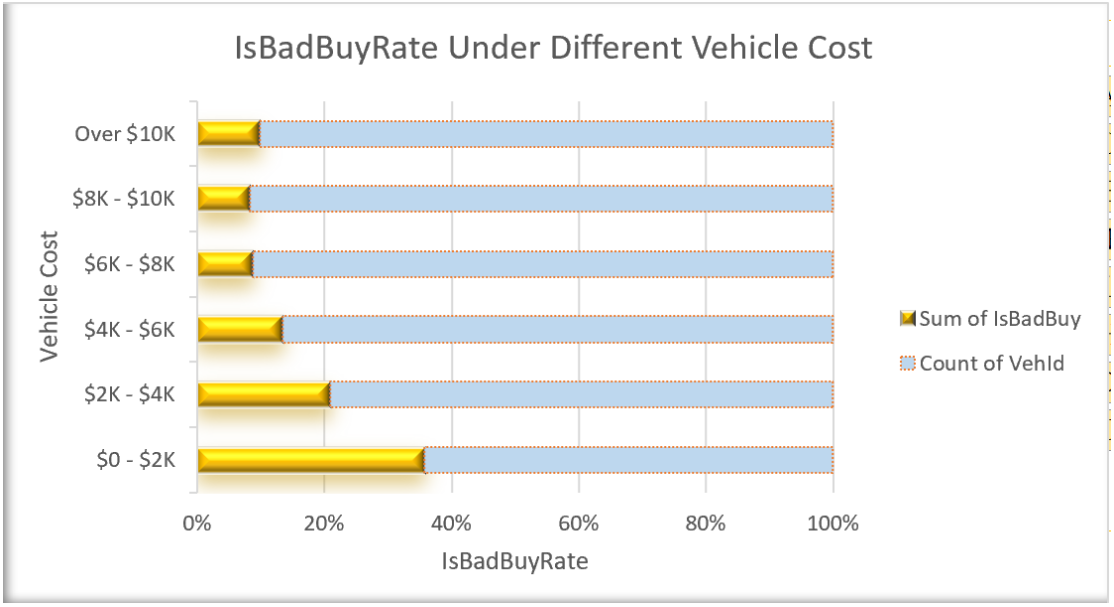
Based on the graph below, we can find out some relationships between IsBadBuyRate and used car odometer readings. The reason I use IsBadBuyRate but not the counts is that each miles range has different counts of BadBuy, so we cannot simply use counts

to compare the relationship between them. Thus, the IsBadBuyRate in every mile range is based on how many total badbuys in that range.

First, we see a lot more bad buys concentrated in the beginning 0-20,000 miles and middle mileage ranges, specifically between 40,000-60,000 miles and 60,000-80,000 miles. This suggests vehicles in this mileage have a higher chance of being labeled a bad buy, maybe because latent defects or wear-and-tear issues start to surface after this level of use.

Also, the visualized data shows lower bad buy rates at the extreme ends of the mileage spectrum. In the 0-20,000 miles range, when vehicles are still relatively new and lightly used, bad buy sums are higher than 60,000-80,000 miles. A possible reason is that it is unusual to have cars that still under 20,000 miles would be sold as used cars unless they encountered serious accidents or malfunctions. Thus, these damaged cars will easier to be a bad buy even with extremely low odometer readings.

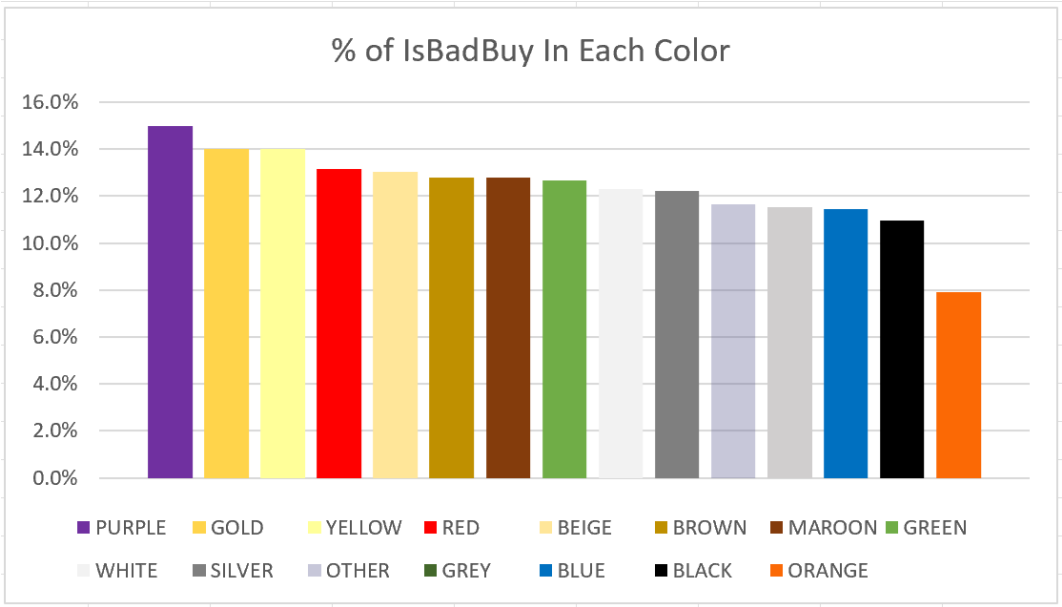
2. IsBadBuyRate Under Different Vehicle Cost



In here, the IsBadBuyRate in every range is also based on how many total badbuys in each range.

Vehicles in the lowest price range ($0 - $2K) have the highest rate of being labeled a bad buy, as indicated by the longest yellow bar. Which is exactly what we assumed, when the sale prices of cars are lower, the situation of those cars are more likely to be bad. As the vehicle cost increases, the IsBadBuy rate seems to decrease, with the Over $10K category having the shortest yellow bar, which suggests a lower rate of bad buys.

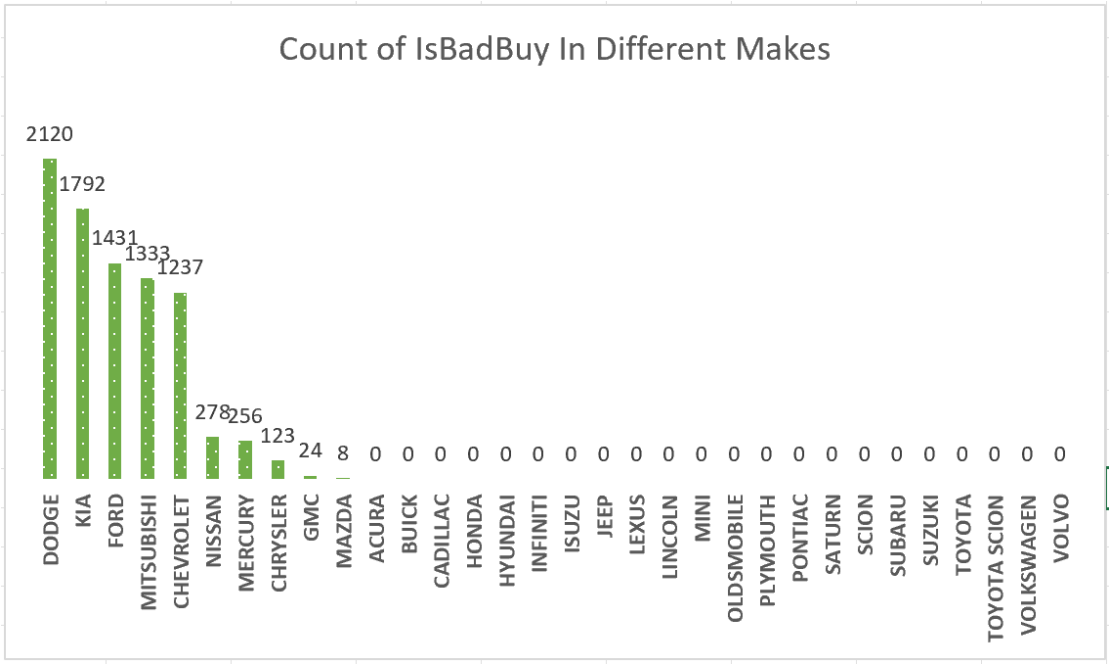
3. IsBadBuyRate Under Different Vehicle Color



In here, the IsBadBuyRate in every range is also based on how many total badbuys in each range. In order to make this graph easier, I made color fit in the color of cars they represented.

In this part of analysis, I was trying to understand if the Vehicle color will be a factor of badbuy. The reason I did it is simply because when I bought my car insurance, I heard an assumption that red cars tend to be charged more insurance fee because they are easier to have car accidents statistically. Surprisingly, As the graph shows below, the purple car is the highest with 15% IsBadBuyRate, which is 3% higher than average.

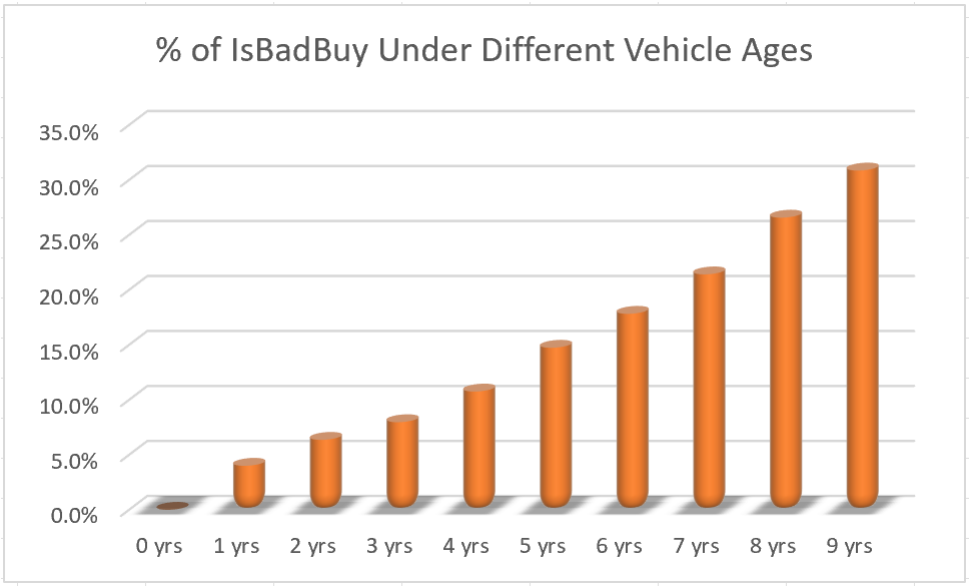
4. Count of IsBadBuy In Different Makes



The chart shows multiple brands, but certain brands (e.g. DODGE, FORD,

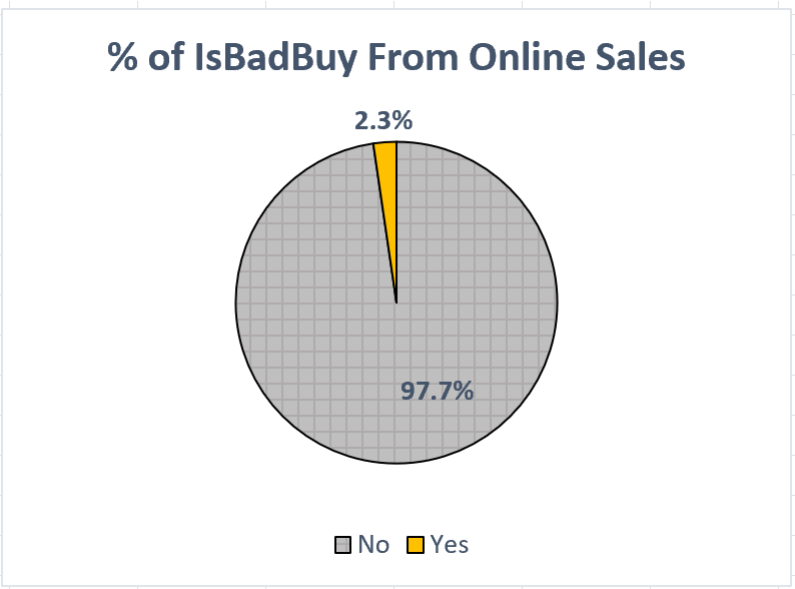
MITSUBISHI, CHEVROLET) have a much higher number of bad buys than others.
The DODGE brand had the highest number of bad buys at 1,792, FORD and MITSUBISHI brands also have high numbers of bad buys, at 1,431 and 1,333 respectively. In contrast, some other brands such as HONDA, HYUNDAI, JEEP and LEXUS have very low bad buy numbers, close to or equal to zero. These numbers could mean that certain brands of vehicles are at higher risk on the used market or may not be as good as other brands in terms of quality.
A vehicle's brand may be one of the most important factors influencing its identification as a bad buy. However, this could also be a possibility that this data is a little bit biased, since this data is gathered in the U.S. so tend to have more U.S. brands' data.

5. IsBadBuyRate Under Different Vehicle Ages



The IsBadBuyRate in every range is also based on how many total badbuys in each range.
The graph shows that the likelihood of a vehicle being a bad buy increases with its age. The lowest percentage of bad buys is associated with vehicles that are new (0 years old). There is a noticeable increase in the percentage starting at vehicles that are 4 years old, with the trend continuing to rise sharply with age. The highest percentage of bad buys is among vehicles that are 9 years old.
This trend suggests that as vehicles age, the risk of them being deemed a bad buy due to potential issues such as wear and tear, outdated technology, or expensive maintenance increases. The graph indicates that the steepest increase in bad buy percentage occurs after the 4-year mark, possibly reflecting the point at which warranties expire and maintenance costs rise.

6. IsBadBuyRate From Online Sales



The IsBadBuyRate in here consider all the badbuy cars in this dataset at the same time, and figure out there is only 2.3% badbuy cars are purchased online. This finding is unexpected for me as I was believing that when the buyer cannot check the used cars in person, it tends to be riskier. However, only 2.3 % badbuy cars are purchased online probably means either online website do a really good job in car quality checks, and provide pretty transparent diagnose for each used cars.