

Auto-Dealership Analysis

Aditya Parashar

The project aims to gather insights and develop use-case specific indicators about Auto dealers using publicly available information

Problem Statement

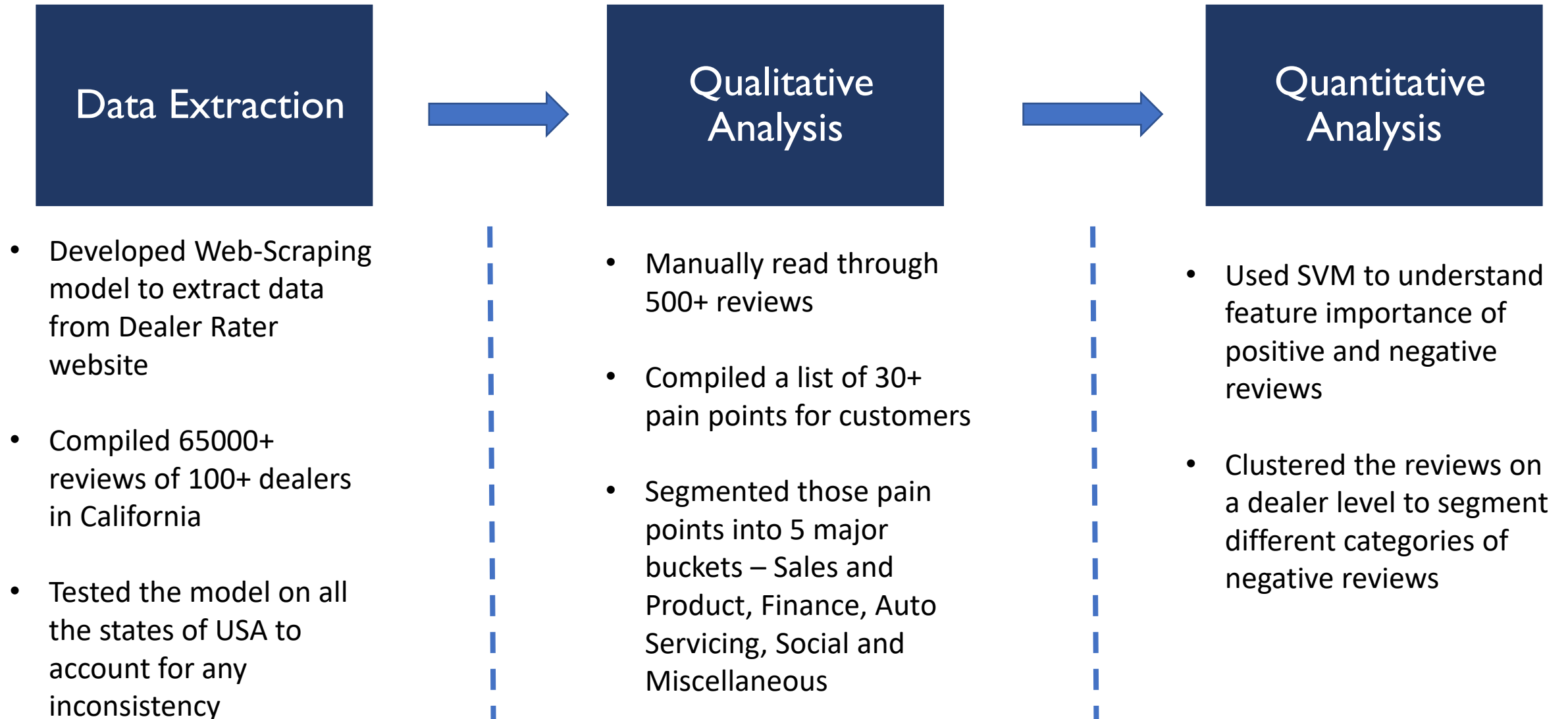
To summarize key information about US based Auto dealers using publicly available information (primarily customer reviews) and to develop models for use cases focused on identifying indicators of customer-centric and dealer-centric issues.
Examples of potential use-cases:

- Dealer is engaged in fraudulent activities
- Current team is using deceptive tactics for increasing sales

Proposed Approach

- Determine and collect relevant online data in usable form
- Conduct qualitative analysis to gain problem specific insights
- Carry out quantitative analysis to understand patterns and underlying structures
- Finalise use-cases for given data considering usability, impact and feasibility
- Construct ML /rule-based models to gain insights for different scenarios

We extracted data from Dealer Rater ,conducted qualitative analysis to understand the data followed by a preliminary quantitative analysis



Agenda

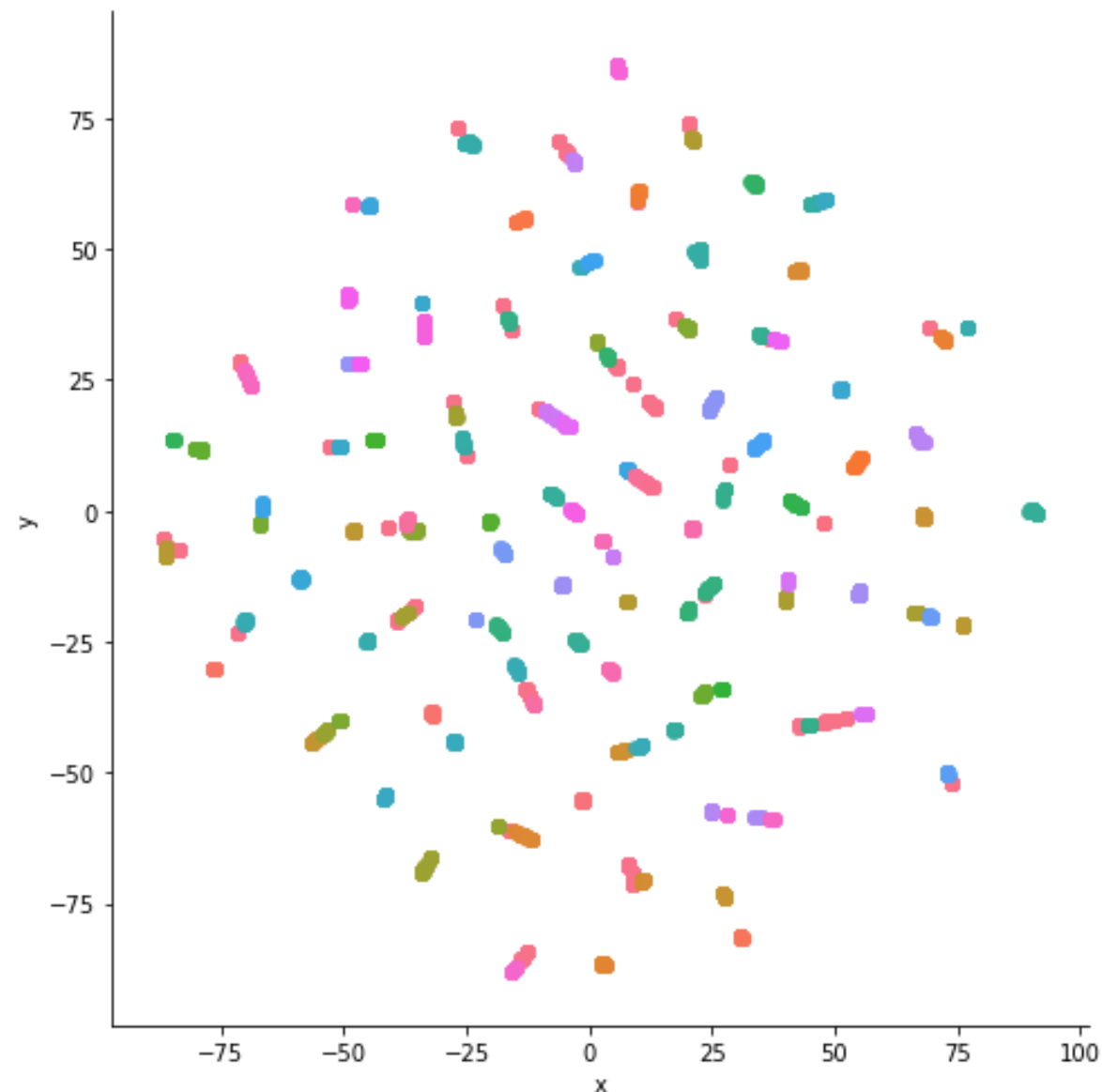
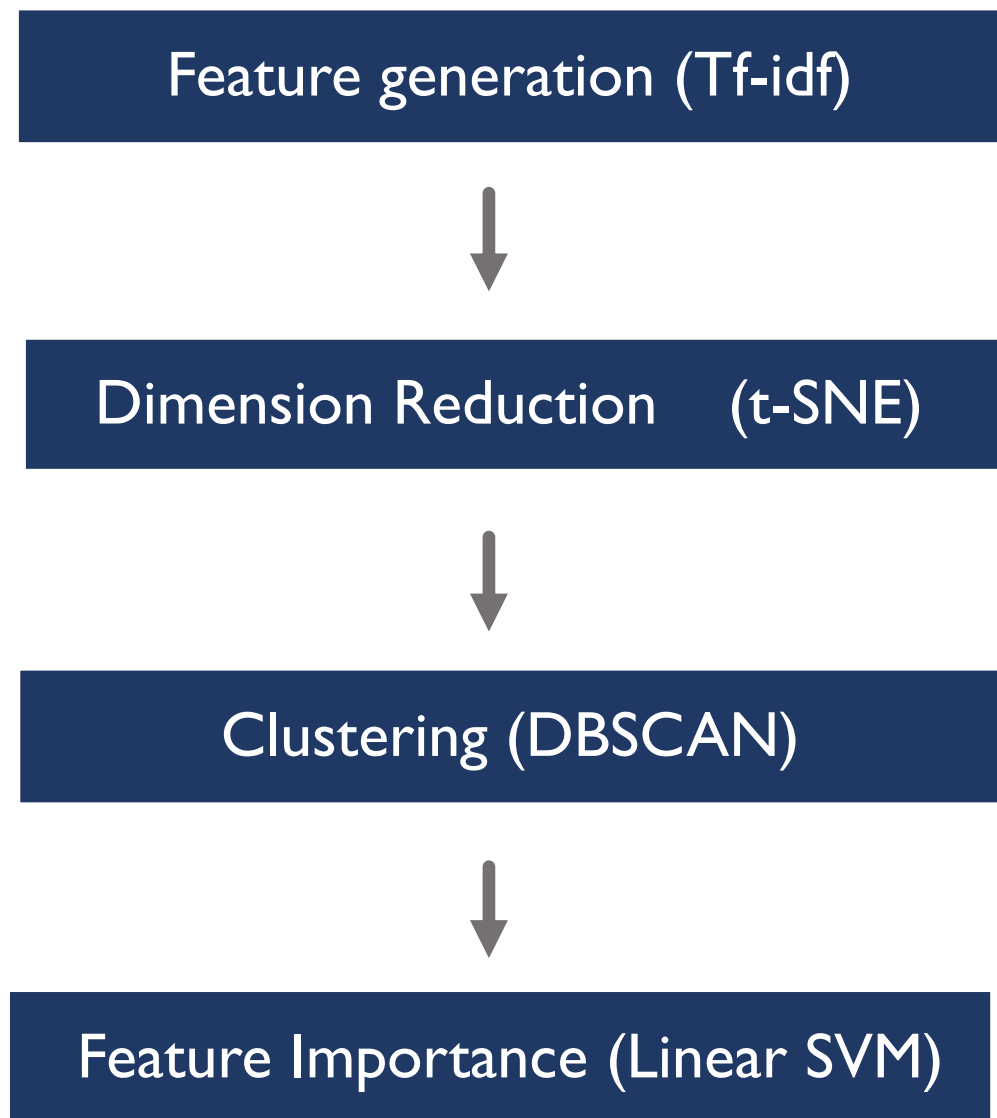
Identifying Use Cases

Data Annotation & Model Building

Insights

Conclusion

We segmented all negative reviews into 110 Clusters using t-SNE and DBSCAN and identified key word-features of all the clusters using Linear SVM



We filtered the segmented clusters on the basis of interpretability, impact and relevance thus arriving at 8 potential use-cases

110 Total Number of Clusters

81 Interpretable

22 Relevant

8 High Impact

Word Features describing the clusters:

- *Advertised, Price*
- *Interest, Loan*
- *Contract, Read*
- *Dishonest, trade*
- *Warranty, Covered*
- *Title, Paid*
- *Credit , Score*
- *Refund, amount*

We finalized five use cases based on qualitative survey and results of clustering analysis of reviews

Potential Use-Cases

➤ **Discrepancy in Advertised Price**

Dealers bring customers in by offering cheap prices online which vary significantly from final offered price

➤ **APR or Loan Issues**

APR is very high even for customers with good credit or interest rate/loan time is increased in final contract without customers' consent

➤ **Spot Delivery Scam**

Dealers allow customers to take their vehicles home and increase APR or down payment later often leading to scraping off of entire deal

➤ **Refund Delay or cancellation**

Dealers do not process or delay the refunds of customers

➤ **Title Issues**

Dealers do not process title registration even after the deal goes through or sell a vehicle of which they don't own the title

● → **Clustering Segment**

● → **Qualitative Analysis**

Agenda

Identifying Use Cases

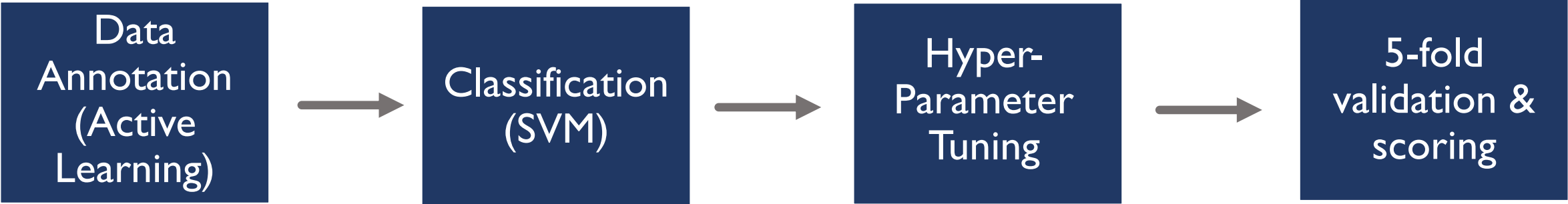
Data Annotation & Model Building

Insights

Conclusion

We built supervised learning model to detect discrepancy in Advertised Prices by using Tf-Idf vectors and SVM with f1 score of 0.89 (California dataset)

Discrepancy in Advertised Price



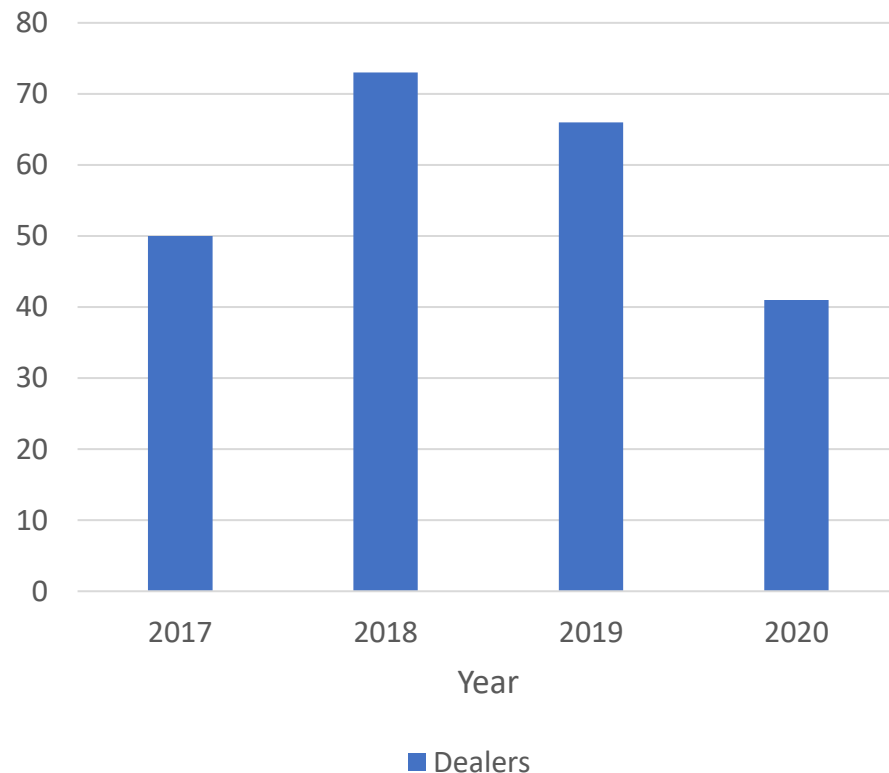
Model Validation & Scoring	Metric	F1-score	Precision	Recall	Accuracy
	Score	0.89	0.892	0.888	0.89

Scores are evaluated using 5-fold validation of California dealers’ dataset (4000 reviews)

We tested the model on Texas dataset (24000+ Reviews, 800+ dealers) and detected 168 unique dealers with false advertisement strategy in 2017-19

Total number of Texas Dealers detected = 168

Precision (Texas Dataset) = 0.8 (Estimated)



Number of Dealers detected year-wise

Examples:

Dealer: West Point Buick GMC, Texas

Upon working the deal Echo attempted to sell the car to me for about \$5000 more than what the advertised price was even AFTER I told him the special. He goes back and comes back and still tries to sell it to me for about \$2000 MORE than what the advertised price was.

Dealer: Maxwell Ford, Texas

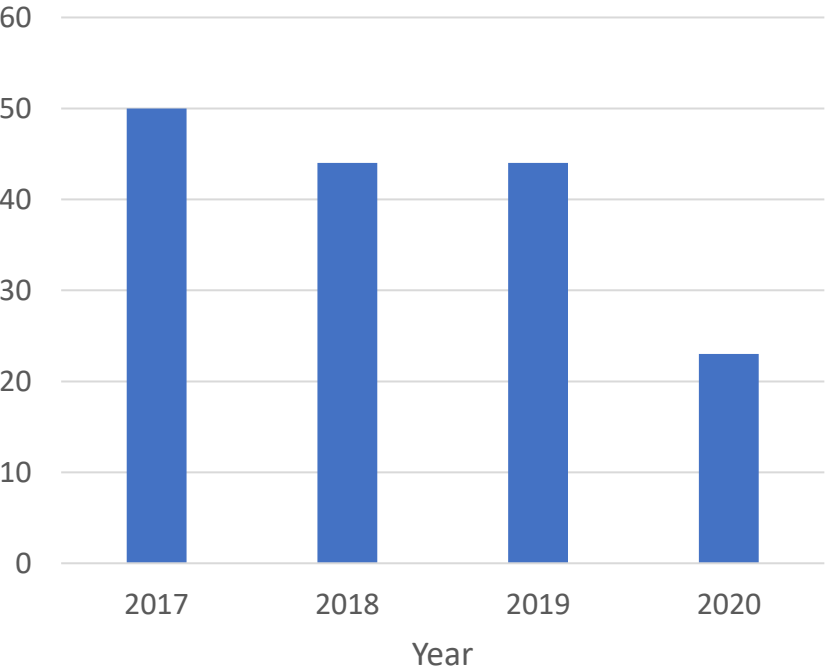
I inquired 2 days ago about a new 2019 Ford Ranger you had listed for \$19460 through auto trader. The salesperson, , Eddie Sifuentes called me immediately. We spoke and I told him I was ready to make the deal. Just needed to know a final price so I can obtain the loan at my bank and get the money. Within an hour or so Eddie calls me. He stated the final price BEFORE taxes, fees, and tag would be over \$27000.

We used similar steps to create and validate model for APR & Loan Issues and tested it on Texas dataset

Model	F1-Score	Precision	Recall	Accuracy
APR & Loan Issues	0.87	0.89	0.868	0.872

Total number of Texas Dealers detected = 130

Precision (Texas Dataset) = 0.78 (Estimated)



Number of Dealers detected year-wise

Scores are evaluated using 5-fold validation of California dealers’ dataset (4000 reviews)

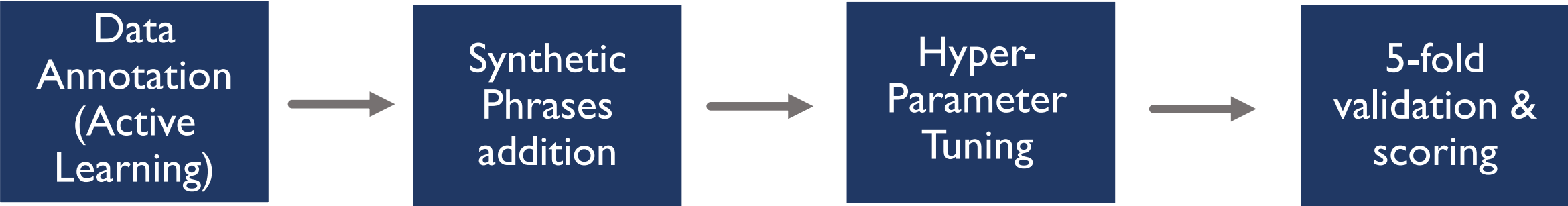
Dealer: Gulfgate Dodge Chrysler Jeep Ram, Texas

I drove 4 hours to come buy 2018 Charger Hellcat with already approved financing from my bank. Upon my arrival the finance manager started to delay the process to where when it was time they needed the transfers instructions the banks were closed and I was being manipulated to where if I wanted to buy the car I had to get the financing through the dealership at their APR %.

Dealer: Big Star Honda, Texas

I do not appreciate the Honda Lender to changed the rate from 3.45 to 3.6% after what we agreed upon. We signed for 3.45%APR for 72mo. We were asked to come back to signed papers they said we missed, we were caught off guard. The accounting said the company did not approved us for 3.45%, make no sense!

Spot Delivery scam showed lower precision than other use-cases. We added phrases showcasing spot delivery scam to increase the performance of the model



Model	F1-Score	Precision	Recall	Accuracy
Spot Delivery Scam	0.806	0.76	0.82	0.83

Spot Delivery Scam had low recall and precision. Possible reason could be more nuanced and detailed nature of this issue than other use-cases.

We tested this model on Texas dataset (24000+ Reviews, 800+ dealers) and detected 22 unique dealers with spot delivery scams in 2017-19

Total number of Texas Dealers detected = 22

Precision (Texas Dataset) = 0.42

Correct Prediction

Dealer: All American Chrysler Dodge Jeep Ram of Midland, Texas

*The next day they said **financing fell through for my father but had another contract "for just a little bit more interest" this is called spot delivery and is done because the dealer gets all of this added interest, and ITS ILLEGAL he didn't have to sign the new contract** but he did. Weeks later driving in my new truck and they call me saying my financing fell through, but I didn't fall for that. We signed a contract if I can't back out when it's signed neither can the dealer. But they have been harrassing me and my father for a month now,*

Incorrect Prediction

Dealer: Maxwell Ford, Texas

*Started out by quoting **a higher price with a higher interest** rate even though we came in with our own financing at a much lower rate. Only dropped their price once we receive the call from a competing dealership for a lower price.*

We built a two-step model- a rule based algorithm followed by supervised learning model to detect title issues and those with discrepancy in refunds.

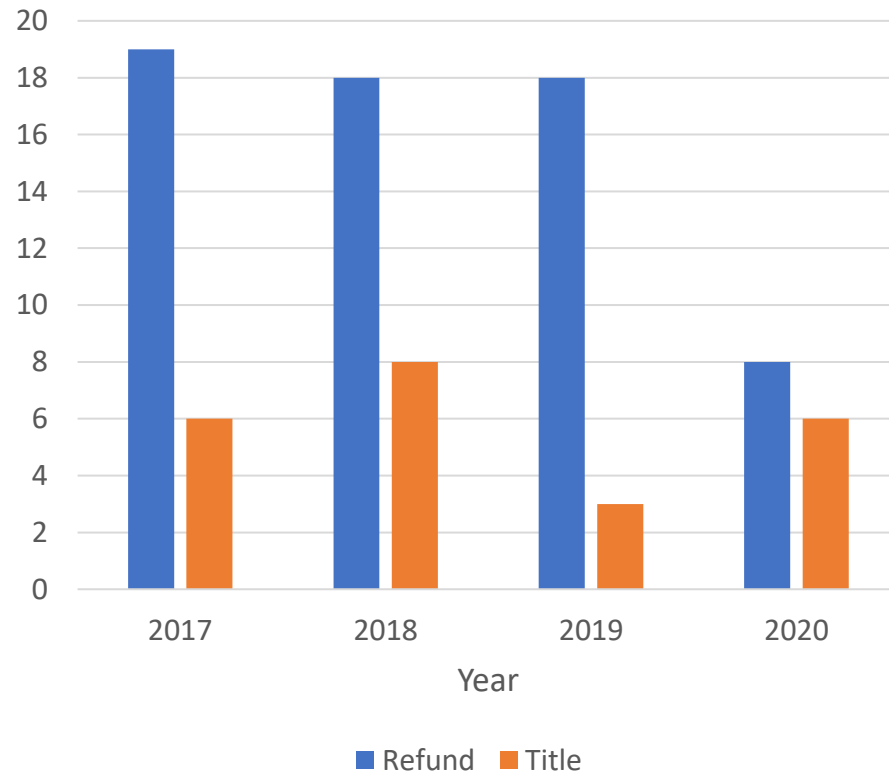


Model Validation & Scoring					
	Model	F1-Score	Precision	Recall	Accuracy
	Title Issues	0.878	0.974	0.832	0.952
	Refund Discrepancy	0.82	0.84	0.77	0.86

We tested models of title and refund issues on Texas dataset and detected 22 & 61 dealers respectively in 2017-19

Precision - Refund (Texas Dataset) = 0.73 (Estimated)

Precision - Title(Texas Dataset) = 0.92 (Estimated)



Number of Dealers detected year-wise

Examples:

Dealer: Stonebriar Chevrolet, Texas

I purchased a used vehicle in October cash. I still do not have plates or a title that say I purchased it. Why would you sale a vehicle that you don't have a title for to be able to sale to someone else? .

Dealer: Joe Myers Mazda Kia, Texas

I have been calling this dealership since June regarding a car that was paid off in April. I have called 7 times and gotten no response. I have not been able to speak to a live individual - calls only go to voicemail. No status on the refund has been received. They have not returned my call. I'm wondering if they will refund the money owed to me..

Agenda

Identifying Use Cases

Data Annotation & Model Building

Insights

Conclusion

We scored the results of Texas Dataset on three factors – Freq. of use case specific reviews of 2019-20, multi-year detection and presence in multiple verticals

Frequency of use-case specific detection

- Dealers were scored based on frequency of detected reviews for each use-case (fraction of total annual reviews)
- Fraction of total reviews was used to take size & popularity of dealer into account

Multi-year detection

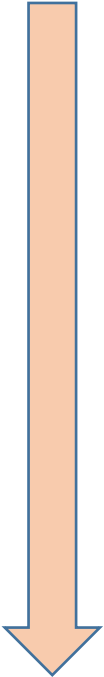
- Dealers with consistent presence (multi-year) were marked
- Dealers with a specific issue for multiple years increases the chance of policy of common malpractice

Presence in multiple verticals

- Dealers detected in multiple verticals were tagged
- Presence in multiple verticals showcases widespread suboptimal practices

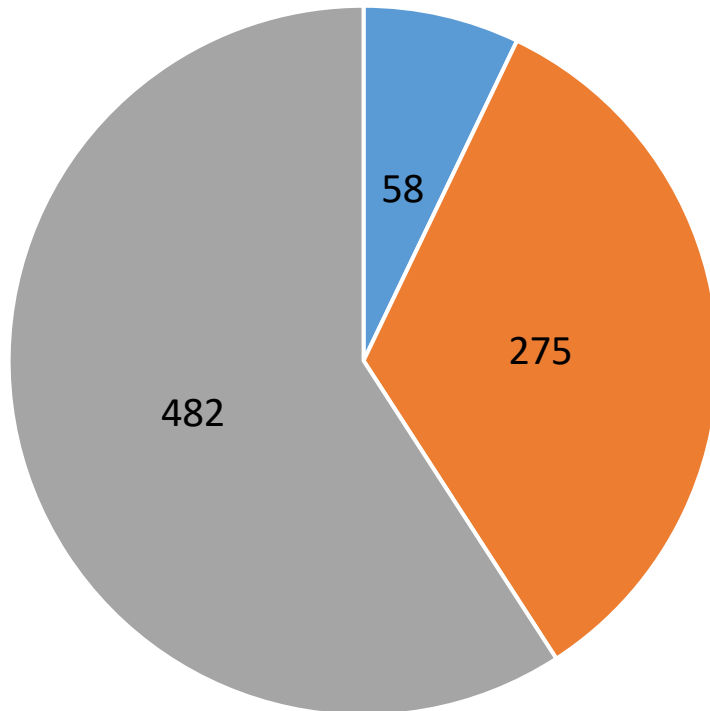
We segmented these dealers into three categories – High Risk, Medium Risk and Low Risk

	Frequency	Multi-year Detection	Presence in Multiple Verticals	Top- Down Classification
High Risk	Top 10%	Last two years (2018-20)	3-4 Verticals	
Medium Risk	10%-30%	Any two years	2 Verticals	
Low Risk	Below 30%	One year	1 Vertical	



- Dealers were segmented in a top-down approach i.e. first high priority dealers ,then medium and finally low to avoid multiple classifications
- Dealers were segmented into these classes if any one of the criteria were satisfied

We created a list of dealers of Texas belonging to different segments



■ High ■ Medium ■ Low

Dealers in Texas

Example of a high-risk dealer:

Tom Peacock Nissan, Texas

- Has negative reviews belonging to three segments – Advertisement, APR and Refund
- Detected in multiple years

Reviews:

- *But that's not what was told to me in the beginning Mike... He tells me that I should just return the car because it was a bad deal and I should have never been able to drive off the lot.*
- *dealer add-on charges (that we were told we did NOT have the option to decline as they were "already installed") such as nitrogen tires, door guards, and GPS??+ additional ~\$1000 or so in total. PLUS the tax/title/license, for a grand total of almost an additional \$5000 over the advertised price of the vehicle.*

Agenda

Identifying Use Cases

Data Annotation & Model Building

Insights

Conclusion

We can use the results of this analysis in Issue-specific , Dealer specific and segment specific use cases

Issue specific use-cases

- Capital One Dealers involved in misinformation on APR and loan terms can be identified for closer monitoring
- Digital Focus can be re-evaluated for the dealers who engage in sub-optimal advertising tactics

Dealer specific use-cases

- Dealer scores can be used to keep track of dealers' performance & annual/quarterly data can be leveraged to identify dealers who have recently engaged in sub-optimal tactics

Segment specific use-cases

- An additional layer of segmentation which includes high risk dealers based on given use-cases can be used to incorporate customer preference and experience in already existing models on dealer risk

Thank You

Appendix

We selected Dealer Rater as the data source for this project

- Dealer rater had dealer specific reviews easily accessible rather than car/model-based reviews generally popular among its competitors
- Location wise segmentation was more organized hence making it easier to develop a web scraping model
- 6 Million+ reviews were available segmented into Sales/Service review for all the dealers
- Strict policy for fraudulent reviews with all the reviews being checked manually before being uploaded to the website

Future scope of this analysis:

Exhaustive

Integrating more use-cases to create more accurate and exhaustive set of models to analyse Dealer risk

More Accurate

Increasing the amount of data used to train and validate the models thus improving its performance

Scaling Up

Scaling the model up to include multiple states and validating the results using already existing set of models

We identified dealers with distinctive issues based on the results (California & Texas)

Examples :

Audi Central Houston

- *5 different instances of price discrepancy in the 4th quarter of 2018*
 - *All of the negative reviews were based on/consisted of huge discrepancy in offered and advertised price*
- E.g. : When I ask for a specific car they showed they had online they showed it to me but the price was about \$10,000 more than what they showed it to be.*

Rusnak Pasadena

- *12+ reviews claiming huge delay or denial of refunds consistent across 2017-20*
- E.g.: Unfortunately, the mistake was never corrected and we had to cancel the order. Although the dealership says they have refunded our deposit, they have not*

Honda World Westminster

- *2 complaints of spot delivery scams within a month*
- E.g. : I went to this dealership to buy a new car. I left with what I thought was a good price and a nice car. A week later I received a letter saying "notice of election to rescind contract."*

Pain points identified during qualitative analysis:

Sales and Product

- *Availability of the advertised car*
- *Condition and Quality of Used/New Car (Often determined by number of days it took for a new/used vehicle to be returned for immediate repair)*
- *Breaching of Contract - (Specially in case of returns , trade-in cars)*
- *Return Policy*
- *Availability of loaner car and/or ride back*
- *Price compared to competitors*
- *Prior info about used cars - (Car previously crashed etc.)*
- *Charging higher than advertised prices (Non-optional additional packages on top of advertised price)*
- *Low valuation of trade-in vehicles*
- *After/During the negotiation , change in availability or price of the vehicle*

Social

- *Rude or Disrespectful Employees*
- *Facilities for customers in dealership (One off)*
- *Differential treatment for vehicles bought from other dealers*

Miscellaneous

- *Covid 19 Response*
- *Fraudulent Activities - (Instances of scamming via credit card/ bank account details, Spot Delivery Scam)*

Finance

- *Ease of obtaining loans and favourable/Unfavourable terms*
- *Delay in refunds*
- *Policies revealed after/at a later stage of a deal (No company cheques, Return and Refund policy etc.)*
- *Dealers' Inclination towards loan financing rather than on-cash deal*
- *Change in leased car policy*
- *Financial decision without prior permission*
- *Spot Delivery Scam*
- *Misinformation about the details of the deals(Price, Loan Amount etc.)*

Auto-Servicing

- *Time taken in after sales services (Specially Regular Service Checks)*
- *Range of areas where After-Sales Services would be subsidized*
- *Quality of Customer Service (Ease of making appointments)*
- *Inventory Mismanagement (Unavailability of parts at the last minute)*
- *Unauthorised non-subsidized services (Not covered under warranty)*
- *Low response of the service department (Designated service and sales advisors)- **Major Issue***
- *Quality of repairs*
- *Unexpected delays or frequent trips required for servicing (Mismatch in duration of service proposed and required)*
- *Unavailability of designated employees even during scheduled appointment*
- *Intentional change in service requirements to include non-warrantied products*

Tf-idf Feature

- *TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.*

t-SNE

- *t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data. In simpler terms, t-SNE gives you a feel or intuition of how the data is arranged in a high-dimensional space. It was developed by Laurens van der Maatens and Geoffrey Hinton in 2008.*

DBSCAN

- ***Density-based spatial clustering of applications with noise (DBSCAN)** is a well-known data clustering algorithm that is commonly used in data mining and machine learning.*
- *Based on a set of points DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.*
- *The DBSCAN algorithm basically requires 2 parameters:*
 - ***eps:** specifies how close points should be to each other to be considered a part of a cluster. It means that if the distance between two points is lower or equal to this value (eps), these points are considered neighbors.*
 - ***minPoints:** the minimum number of points to form a dense region. For example, if we set the minPoints parameter as 5, then we need at least 5 points to form a dense region.*