SpringBoard Capstone 1: An Analysis of Data from Beepi Inc.

Introduction

Beepi would have been my main client since the data for my capstone project primarily comes from their business. Beepi was a startup in based in Mountain View, CA whose mission was to improve the state of the pre-owned auto market in the United States by offering customers an unparalleled selling or purchase experience. The company was unique in the auto-sales space because it had no physical location customers could visit; instead, customers interacted with the company primarily through the Internet and via telephone. Sellers on the Beepi platform would first need to guarantee that their vehicle was less than 6 years old and had less than 80,000 miles on the odometer. Then, Beepi would send an inspector to review the vehicle and to take photos of it for its listing on the website.

To reassure Buyers, Beepi offered customers who purchased a vehicle from the company a 10-day/1,000 mile money-back guarantee – buyers would love their car or they could return it for a full refund. In a similar vein, sellers would get a guaranteed price at which their car was worth and was guaranteed to sell within 30-days or Beepi would purchase their vehicle at the guaranteed price. All prices that Beepi gave were supposed to beat industry standards by offering more to sellers and providing buyers with a better deal. Therefore, pricing became a key business component that needed to be managed with care since it touched the heart of the company's purpose. Customers who sold their vehicles to Beepi wanted to maximize the value of their asset and customers who purchased a vehicle from the company wanted a fair deal for a high-quality car. Beepi had to manage these pricing expectations all while producing a reasonable sales commission that contributed to revenue.

Unfortunately, Beepi did not succeed in its mission and the company ceased operations in January 2017 after a failing to raise a round of capital. One of the primary contributing factors in the company's demise was vehicle pricing; the company was unable to manage its unit economics and failed to produce profits or show a reasonable path to profitability. My report will focus on why the company failed by examining the vehicle transaction data. I will attempt to identify Beepi's business model strengths and weaknesses as well as what could have been done to potentially avert a company failure. My final portion of analysis will focus on building a pricing model that Beepi could have implemented to offer customers pricing on their vehicles.

Before I continue with my report, I should note that pre-owned vehicle pricing is a tricky blend of art and science that no model can fully capture without unrealistically high levels of detailed data. The main source of this difficulty comes from the fact that each used vehicle on the road is unique; no two cars are the same. Each owner has driven their car differently according to their driving style over different roads and in varying traffic conditions. Each vehicle also has its own unique service history and how it was cared for inside and out.

The Main Focus of Study

Since Beepi's fate as a defunct startup is already known, I plan on doing an autopsy of the company using its data. First, I will explore the dataset and focus on the empirical reasons why I believe the company ultimately failed. In light of these findings, I will make recommendations as to what could have been done differently to correct the business weaknesses found. Finally, I will attempt to build a predictive model for vehicle pricing.

This last component will be by far the most difficult challenge out of the problems stated here. As mentioned above, used vehicle valuation is determined by a host of factors that are not often recorded as data points. For instance: paint quality, tire condition, and the condition of the interior are all major contributing factors to a used vehicle's price which are rarely recorded for analysis after a sale has been completed. I will attempt to do my best to create a reasonable vehicle-pricing model based on my knowledge of the auto industry and with pre-owned vehicle pricing.

The Beepi Dataset

The complete Beepi dataset is comprised of 3 CSV files:

1. 'AllCars.csv'
2. 'Beepi_prices.csv'
3. 'Edmunds_MakeModelStyles.csv'

A supplementary dataset was used for linear regression analysis of vehicle prices:

1. 'true_car_listings.csv'

The first dataset, 'AllCars.csv', does not include every vehicle entered into Beepi's database; rather it focuses primarily on vehicles that were signed up for an inspection, meaning the seller made it past the first "hurdle" in the sales process. However, it includes highly detailed information about each vehicle that entered Beepi's pipeline. For example, for vehicles that were successfully listed on the website, it includes the number of Internet views that vehicle received.

The second dataset, 'Beepi_prices.csv', is much larger dataset that encompasses all vehicles entered into the Beepi database regardless of whether a seller scheduled an inspection for their vehicle. This resulted in many empty/null values since the barrier to entry was very low. Furthermore, while it is larger than 'AllCars.csv', it does not have the same level of detail included with 'AllCars.csv'.

The third dataset is Beepi's mapping tool which allows users to match a particular vehicle with a unique "StyleID" that identifies a vehicle with fine, but not 100%, detail. For instance, although the specific trim level of each vehicle is captured, it does not reveal any information about specific options that can be ordered with each vehicle (e.g. standalone navigation option, leather interior, etc.).

A major hurdle that I had to overcome was the paucity of data the Beepi datasets have at the trim level. While there are several thousand transactions within the datasets, once I pare down the data from make and model down to trim, there are perhaps only 10 to 20 data points to work with. Seeing this as unacceptable, I sought out another dataset and found one on Kaggle which included over 850,000 used vehicle transactions scraped from TrueCar that could augment my own datasets.

Cleaning the Beepi dataset was very difficult and comprised much of my work. There were many blank, null and NaN values that need to be stripped and/or converted to useable formats for analysis. For example, just importing 'AllCars.csv' proved to be a challenge; I needed to use a specific encoding (ISO-8859-1) to read the file into a Pandas dataframe. Fortunately, the Kaggle / TrueCar dataset required minimal cleaning before being usable.

Other Datasets That Could Have Been Used

There are other vehicle datasets available on websites such as Kaggle that target pricing as a main goal/mission. Unfortunately these datasets were not compatible with my capstone for one or more of the following reasons:

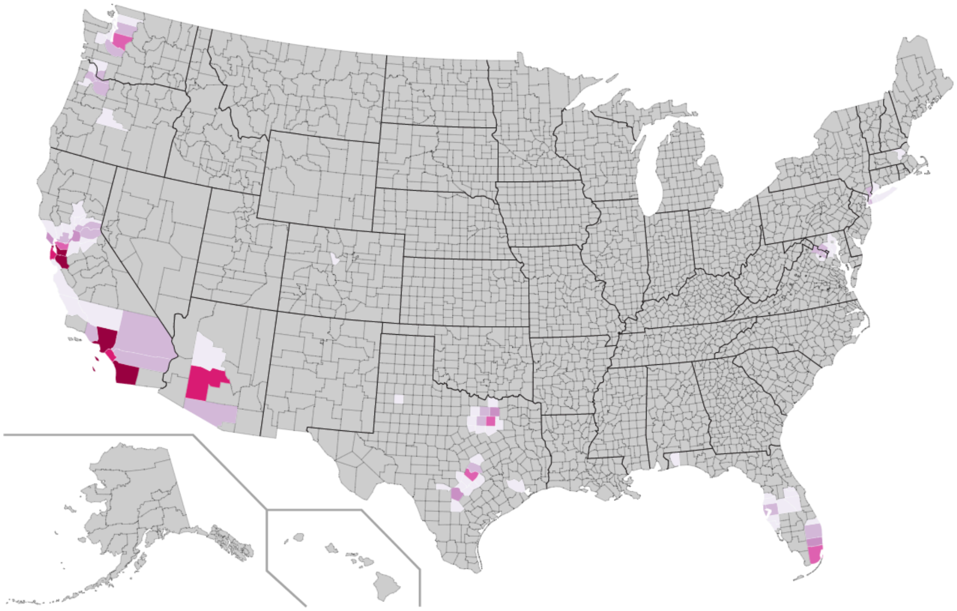
1. The data comes from outside of the United States market (e.g. European market vehicles)
2. There is not enough granularity in the data provided (e.g. make and model are included, but no trim level is provided about the vehicles in the dataset)

A rich source of potential data would be eBay Motors. However, this would require advanced data scraping capabilities and the ability to store millions of detailed transactions in a data warehouse for parsing and analysis.

Findings

Vehicle Sales Patterns

*Figure 1: Map of Total Beepi Sales as of June 2016**



** Darker colors indicate more sales; lighter color indicates fewer sales; no color indicates no sales*

I am leading off with this graphic to demonstrate the limited reach of Beepi in the United States and the relative paucity of internal data the company had (3511 transactions in all). This map of United States counties shows where Beepi generated sales; the darker the color, the more sales generated. We can clearly see that sales were concentrated in California, particularly in the San Francisco Bay Area (Alameda, San Francisco, San Mateo, and Santa Clara counties) and in Southern California (Los Angeles and San Diego counties). Beepi had a small presence in Texas, Arizona, and Florida, but their combined sales all fell well short of those made in California. Due to the limited scope of this data, I supplemented it with another dataset, which was used primarily in comparing different linear regression models of different vehicle segments.

Time Series Analysis of Sales: Day of the Week

Analyzing the Beepi data, I found that the majority of sales occurred during the workweek (Monday – Friday), peaking on Fridays. This result was surprising because most auto sales occur on weekends rather than during the workweek, since buyers have more free time to visit and spend time at physical dealerships. Beepi's sales model did not include a physical location; all purchases were conducted online through its website or mobile application. This pattern of sales indicates that buyers who utilized Beepi fundamentally changed the way they spent their time searching for and buying a vehicle. This information could be highly useful to a marketing team in terms of when/where to spend advertisement dollars and/or how marketing campaigns should be designed in order to generate the greatest number of sales.

Figure 2: Box Plot of Average Sales by Day

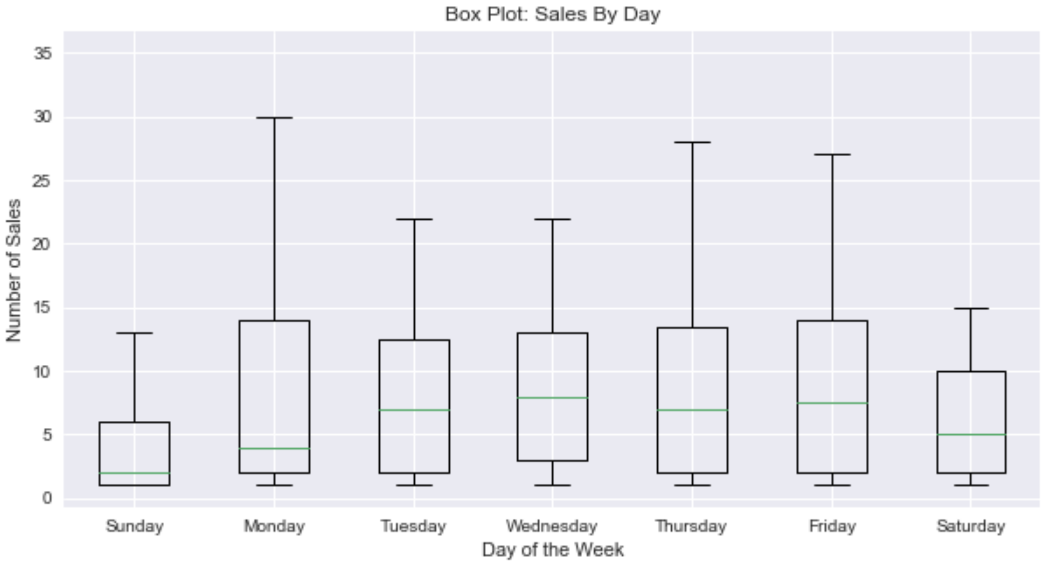
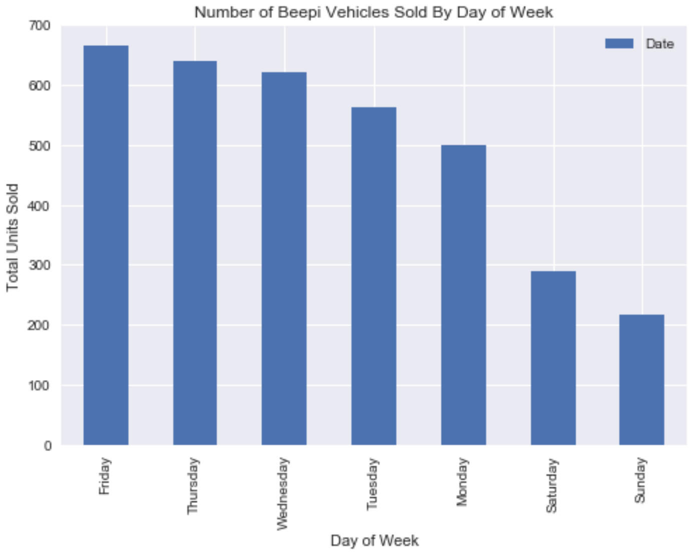


Figure 3: Bar Chart of Total Sales by Day



Time Series Analysis of Sales: Month of the Year

I also examined Beepi's sales performance by month (Fig . It appears that sales in the second half of the year (July – December) were much stronger than sales in the first half (January – June). However, this data must be review cautiously. Company growth, sales campaigns/promotions, and changes to the advertisement stream all contributed to differences in monthly sales. For example, figure 6 shows that sales were depressed for the first 9 months of operations as the company revealed itself to the public. Due to the relatively short operational history of the company, this likely depressed sales for these months and is reflected in the box plot and bar chart.

Without a larger dataset to control for these anomalies, it is hard to determine if higher sales figures in the fall/winter months was due to sales campaigns, buyer habits, or simply growth/more consumer awareness of Beepi as an option when purchasing a used vehicle. Therefore, this data is more useful as a company history rather a platform to derive insights from (March 2014 – January 2017; less than 3 years in operation).

Figure 4: Box Plot of Average Sales by Month

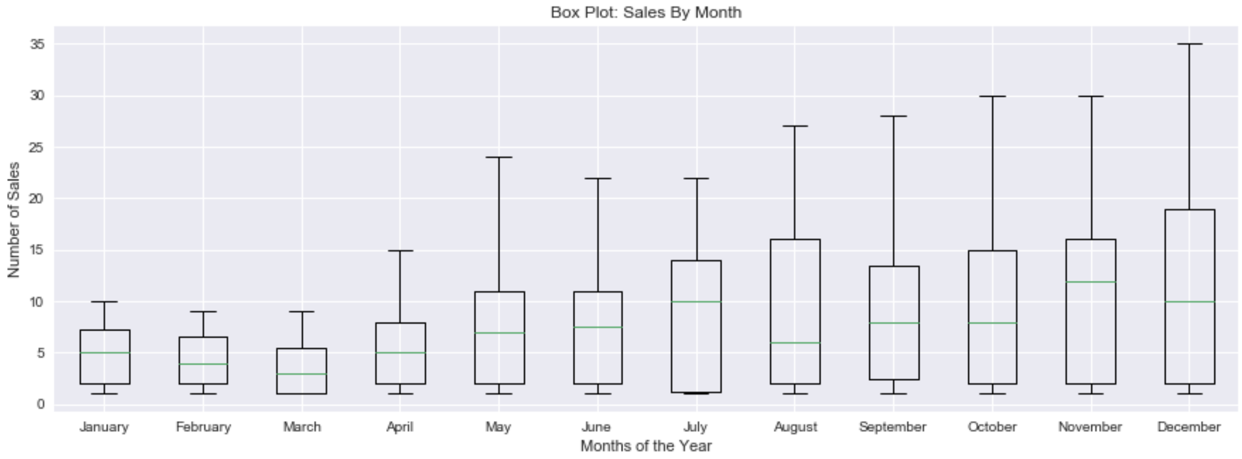


Figure 5: Bar Chart of Total Sales by Month

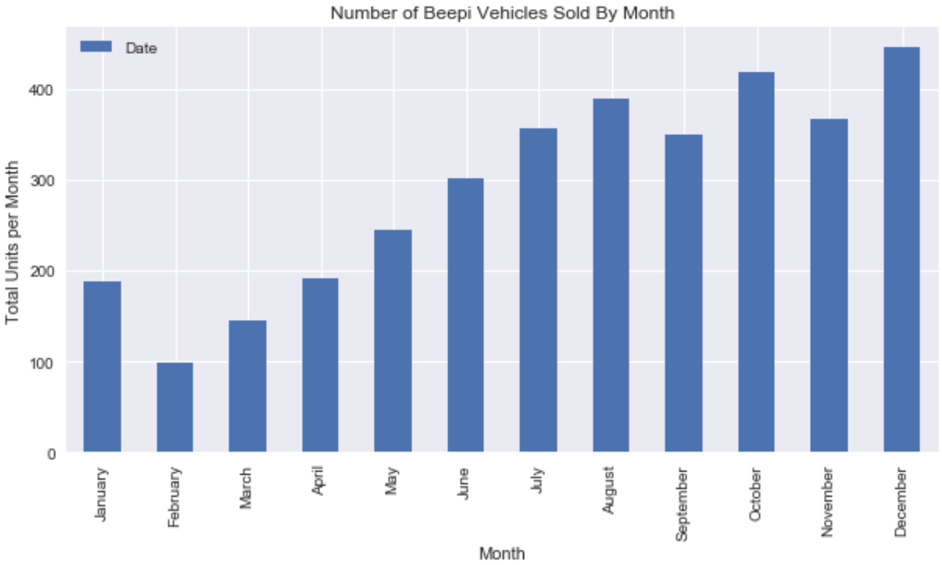
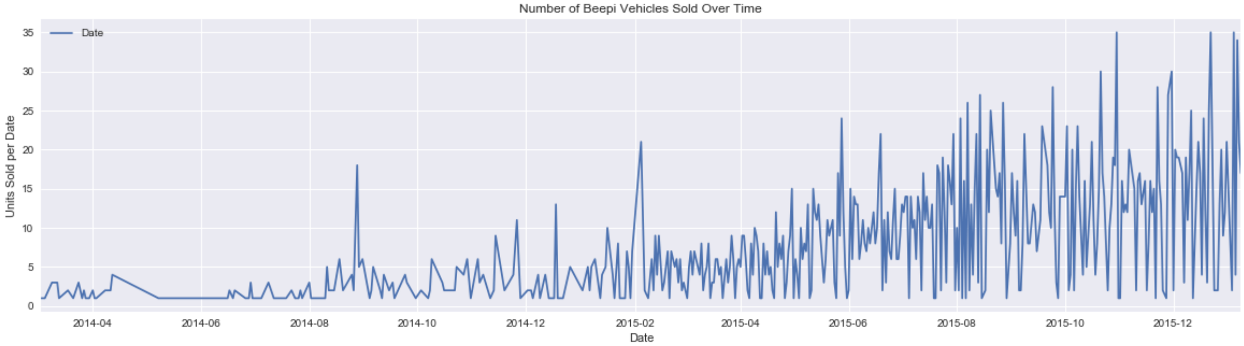


Figure 6: Time Series Plot of all Operational Sales Data



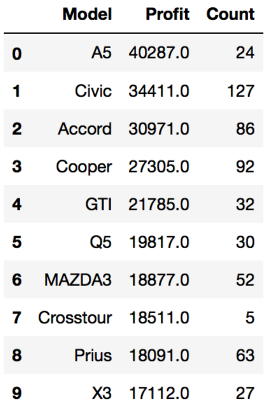
Models with the Highest Gross Profit Margins

Identifying the models that are generating the highest profit margins would be extremely helpful from a marketing perspective. Vehicles that generate interest from buyers could be prioritized in the sales funnel by offering preferential pricing and other benefits to entice sellers to use the Beepi platform and thereby generate more sales and profits for the company. Initially, I reviewed this information purely on a gross-profit basis (Figure 10), but I decided this information would be of limited use to the company's overall profits if we do not understand the per-unit-basis of profit generation.

In order to break this down further, I created two more DataFrames, each of which examine profit generation on a per-unit-basis. Figure ### shows models that are generating the highest profits regardless of unit count and Figure ### shows models that have sold more than 10 units. These findings could be useful for targeting specific sellers with advertisement campaigns or by understanding the preferences buyers have for different vehicles.

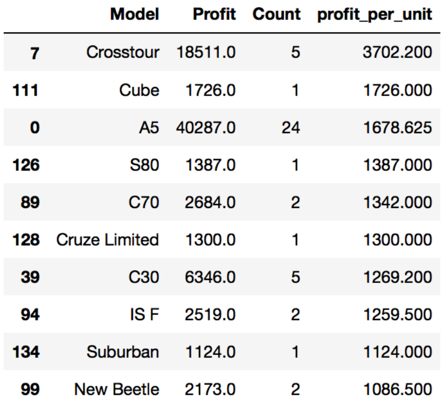
Ideally, I would have coupled this information with the number of days it took to sell each model. One hypothesis I could have tested was: perhaps models that generate the highest profits also sell the most quickly. Unfortunately, the data in Beepi's set is not clean enough to run an analysis of this type since many of the vehicles in the database were sold, returned, and re-sold without keeping track of the sales history; each time a return happened, it overwrote the old data. Due to this data pollution, I elected not to analyze this aspect.

Figure 10



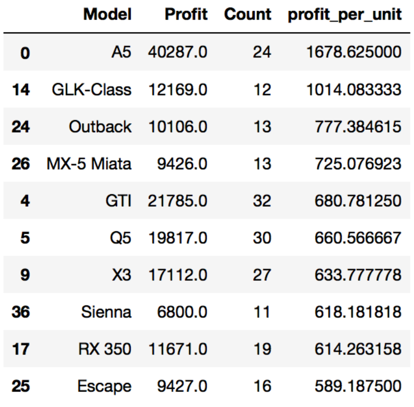
	Model	Profit	Count
0	A5	40287.0	24
1	Civic	34411.0	127
2	Accord	30971.0	86
3	Cooper	27305.0	92
4	GTI	21785.0	32
5	Q5	19817.0	30
6	MAZDA3	18877.0	52
7	Crosstour	18511.0	5
8	Prius	18091.0	63
9	X3	17112.0	27

Figure 11



	Model	Profit	Count	profit_per_unit
7	Crosstour	18511.0	5	3702.200
111	Cube	1726.0	1	1726.000
0	A5	40287.0	24	1678.625
126	S80	1387.0	1	1387.000
89	C70	2684.0	2	1342.000
128	Cruze Limited	1300.0	1	1300.000
39	C30	6346.0	5	1269.200
94	IS F	2519.0	2	1259.500
134	Suburban	1124.0	1	1124.000
99	New Beetle	2173.0	2	1086.500

Figure 12



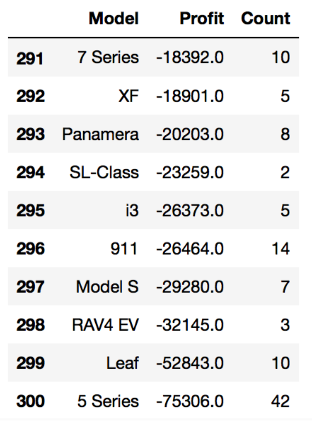
	Model	Profit	Count	profit_per_unit
0	A5	40287.0	24	1678.625000
14	GLK-Class	12169.0	12	1014.083333
24	Outback	10106.0	13	777.384615
26	MX-5 Miata	9426.0	13	725.076923
4	GTI	21785.0	32	680.781250
5	Q5	19817.0	30	660.566667
9	X3	17112.0	27	633.777778
36	Sienna	6800.0	11	618.181818
17	RX 350	11671.0	19	614.263158
25	Escape	9427.0	16	589.187500

Models with the Lowest Gross Profit Margins

I also elected to analyze models with the lowest profit margins, the logic being that this could potentially help stem the losses that Beepi was incurring while it was operating. One immediate aspect that jumped out at me was the number of electric vehicles (EVs) that ended up in part of my analysis. Pricing models at the time Beepi was operating *did not* account for federal and state tax incentives on many EVs, which ended up damaging the company's profits. For example, a new Nissan Leaf cost approximately $30,000, but buyers in California essentially got a $10,000 rebate by combining federal and state tax incentives. Without taking into account these incentives, Beepi was losing money on every EV transaction. My recommendation to Beepi would be to re-adjust internal pricing models to better reflect the EV market or to stop them from entering the marketplace.

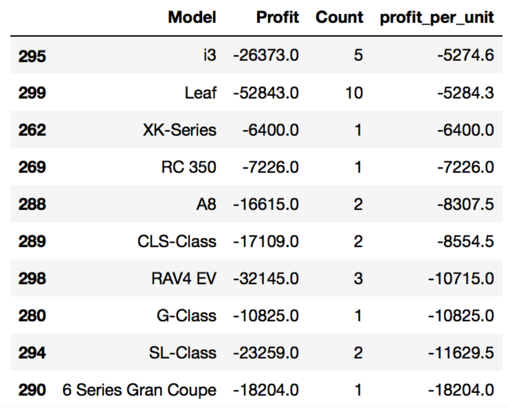
In addition to problems with EVs, we also see that most of the other losses being generated by company come from luxury vehicles such as the BMW 5-Series and the Porsche 911. I would recommend that Beepi revise the pricing models that are used to value these vehicles to better reflect their current market value.

Figure 13



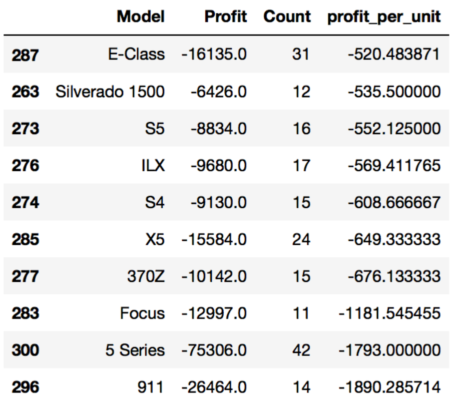
	Model	Profit	Count
291	7 Series	-18392.0	10
292	XF	-18901.0	5
293	Panamera	-20203.0	8
294	SL-Class	-23259.0	2
295	i3	-26373.0	5
296	911	-26464.0	14
297	Model S	-29280.0	7
298	RAV4 EV	-32145.0	3
299	Leaf	-52843.0	10
300	5 Series	-75306.0	42

Figure 14



	Model	Profit	Count	profit_per_unit
295	i3	-26373.0	5	-5274.6
299	Leaf	-52843.0	10	-5284.3
262	XK-Series	-6400.0	1	-6400.0
269	RC 350	-7226.0	1	-7226.0
288	A8	-16615.0	2	-8307.5
289	CLS-Class	-17109.0	2	-8554.5
298	RAV4 EV	-32145.0	3	-10715.0
280	G-Class	-10825.0	1	-10825.0
294	SL-Class	-23259.0	2	-11629.5
290	6 Series Gran Coupe	-18204.0	1	-18204.0

Figure 15



	Model	Profit	Count	profit_per_unit
287	E-Class	-16135.0	31	-520.483871
263	Silverado 1500	-6426.0	12	-535.500000
273	S5	-8834.0	16	-552.125000
276	ILX	-9680.0	17	-569.411765
274	S4	-9130.0	15	-608.666667
285	X5	-15584.0	24	-649.333333
277	370Z	-10142.0	15	-676.133333
283	Focus	-12997.0	11	-1181.545455
300	5 Series	-75306.0	42	-1793.000000
296	911	-26464.0	14	-1890.285714

The Problem with Used Vehicle Valuation

Unlike new vehicle pricing, the difficulty of used vehicle pricing stems from the fact that each one is a unique item: no two used vehicles are exactly the same due to different ownership histories, service histories, and the manner in which they were used. These data points are rarely captured in pricing data and have supposedly made used vehicle valuation a kind of "black art" that only those familiar with the industry can do with a high degree of accuracy. To be fair, these "black art" factors do affect price, but the most important factor in determining price, aside from a vehicle's year, make, model, and trim is factory options selected at the time of purchase.

New car buyers spend hundreds or even thousands (or in extreme cases, tens of thousands) of dollars to have optional extras on top what comes with their vehicle as standard. For example, there are over 400 factory options for the 2016 Porsche 911 Carrera that can add over $50,000 to the final factor MSRP. A 2016 Porsche 911 Carrera with *all* options selected should obviously be valued differently than one with no options selected. Unfortunately, the vast majority of datasets available neglect options and only capture four features that are typically evaluated to price a used vehicle: age, model, trim level, and mileage.

I will attempt to show that options are a driving factor in befuddling many pricing models using two comparisons:

1. 2014 Honda Pilot EX-L v. 2014 BMW X5 xDrive35i
2. 2014 Honda Pilot EX-L v. 2014 Honda Pilot LX

For my first comparison, I have chosen to work with a Honda Pilot precisely because Honda is unique in the automotive landscape in that it sets all options by trim level (in this case, the EX-L trim level). This means that there are no standalone option choices and all vehicles with a designated trim level have the same exact option set. Once we isolate a specific year, model, and trim for any given Honda vehicle, we have controlled our data for options: all of the vehicles in the set have the same options set.

On the other hand, the BMW X5 does not isolate options based on trim level. In fact, once X5 buyers choose a specific trim level, BMW encourages them to add thousands of dollars in optional extras to their vehicle. Leather seating, a premium sound system, and high-performance LED headlights are examples of options some buyers may choose to outfit their X5 with. In the world of BMW, trim level solely designates a model's engine and drivetrain designation; in this case, xDrive35i stands for a 3.0L turbocharged inline-6 motor paired to an all-wheel drive drivetrain.

The Seaborn JointPlots below demonstrate that there is much more price variation in BMW X5s than Honda Pilots. The kernel density plots show that there is a much tighter concentration of data points for Pilot EX-L models versus its BMW counterpart. This is also reflected in the bar charts along the axes of each JointPlot, where we see sharper peaks for Pilot models and more flattened distributions for the X5.

Figure 16: Honda Pilot EX-L JointPlot

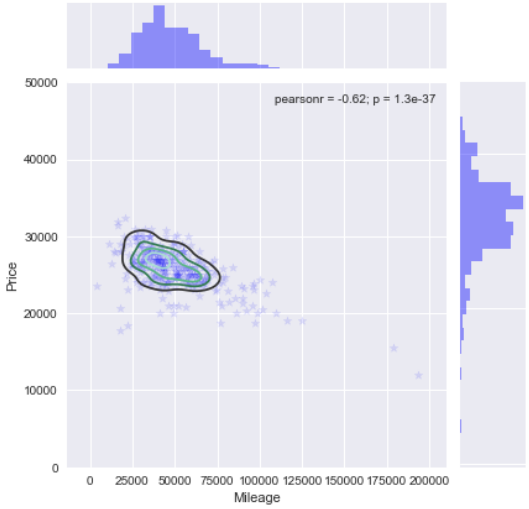
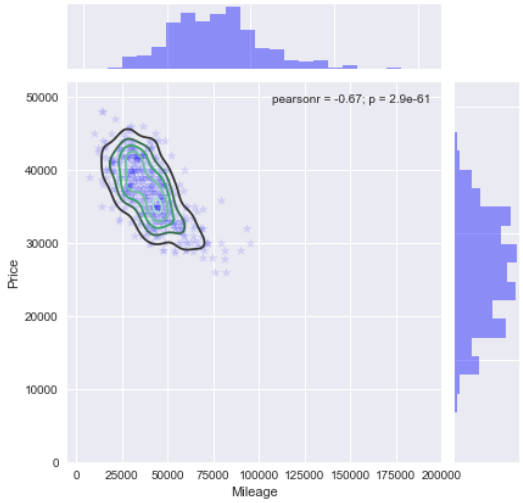


Figure 17: BMW X5 xDrive35i JointPlot



In order to further highlight the effect that factory options play on used vehicle pricing, I will compare the 2014 Honda Pilot EX-L with its “base” model, the 2014 Honda Pilot LX. The Pilot EX-L and Pilot LX are identical vehicles save for the fact that they have different options packages. The EX-L models have niceties such as leather seating, a premium sound system, navigation system, and power-lifting tailgate that the LX models all do without. We can see that the pricing of the Pilot LX model is more linear in shaped compared with the Pilot EX-L, indicating less variance in price. This indicates that used vehicle buyers view options differently; some place more emphasis on their perceived value, while others do not. This difference in view does not happen with the LX model because there are no factory options to evaluate.

Figure 18: Honda Pilot EX-L JointPlot

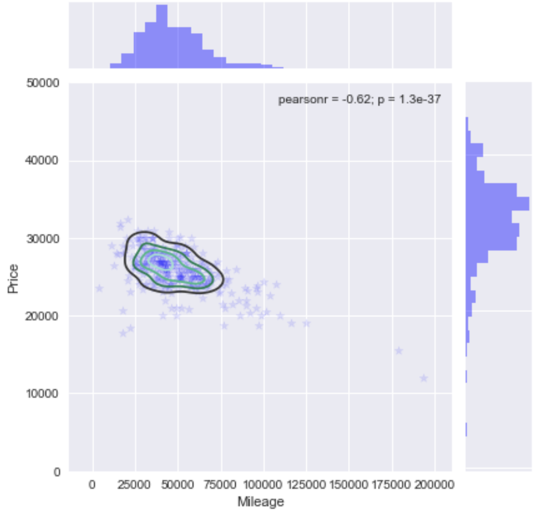
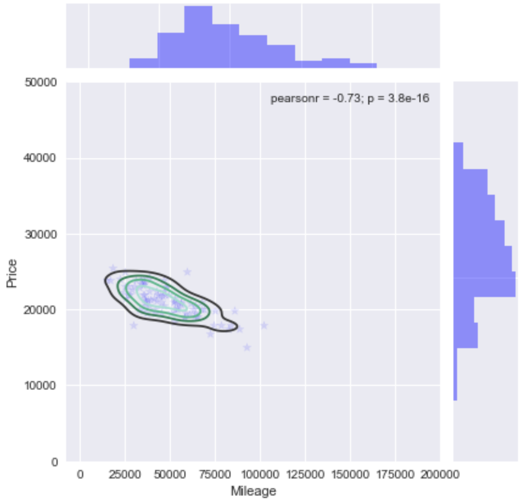


Figure 19: Honda Pilot LX JointPlot



Machine Learning: An Analysis of Linear Regression Models

All vehicles depreciate according to exponential decay. This would be extremely difficult to model if we were attempting to extrapolate pricing into the future based on the data we have, but since we are simply trying to create a pricing model to estimate prices in the present, a linear model will suffice. By controlling for as many factors as possible (e.g. year, make, model, trim level), we can obtain a “snapshot” of the current marketplace of any particular vehicle using a linear regression model to demonstrate how mileage, features, and vehicle quality create variation in the marketplace.

After generating linear regression models for each model and trim level, we can clearly see a stepped difference between the X5 and the Pilot EX-L/Pilot LX models in terms of the root mean square error. The root mean square error was chosen as the evaluator because it represents the sample standard deviation of the difference between the predicted values and the observed values. The X5 has the highest root mean square error (or the highest variance) because we cannot isolate its options based on trim. The Pilot models have smaller root mean squared error values, meaning that they have relatively less variance when compared with models with many options like the X5.

Ultimately these linear regression models can be used as guides to vehicle pricing and each can generate a ballpark estimate that can be shown to customers who wish to sell their car with Beepi. Naturally, a vehicle like the BMW X5 that has many factory options permutations that cannot be easily isolated will have more variance than a vehicle like the Honda Pilot that has factory options that are defined by trim level. By continuously updating these linear regression models with the latest market data, Beepi could have created a valuable valuation tool for its business.

Vehicle	rMSE (root mean squared error)
2014 BMW X5 xDrive35i	$3266.90
2014 Honda Pilot EX-L	$2038.91
2014 Honda Pilot LX	$1510.37

Figure 20: BMW X5 Linear Regression

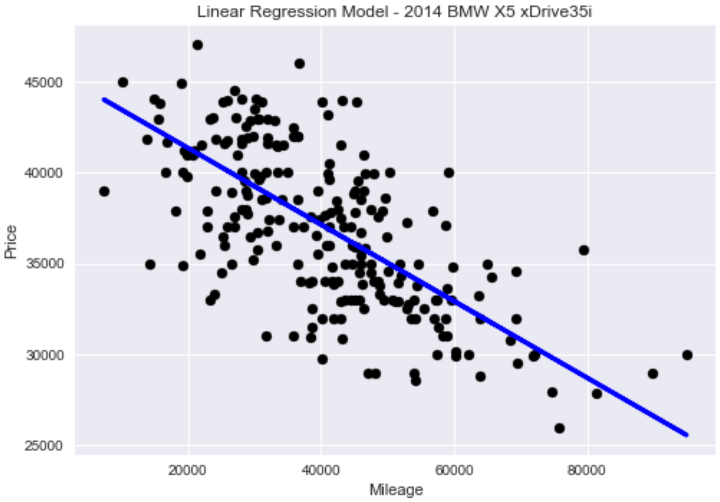
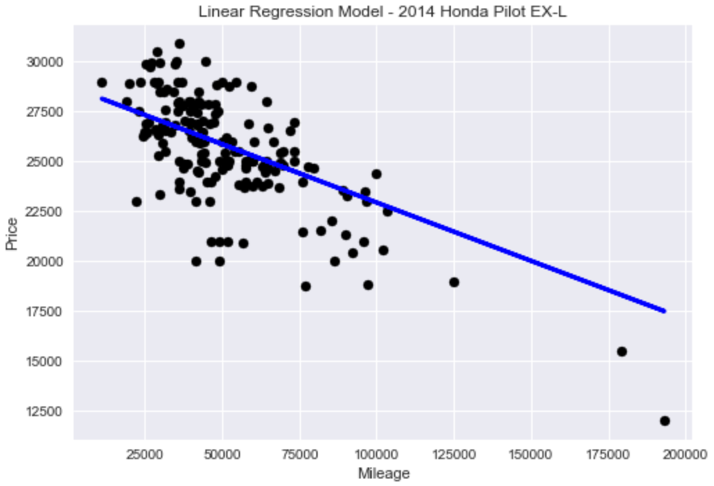
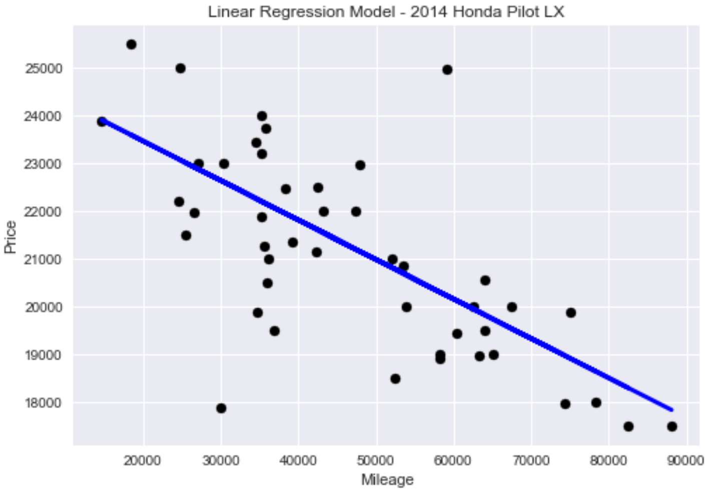


Figure 21: Honda Pilot EX-L Linear Regression

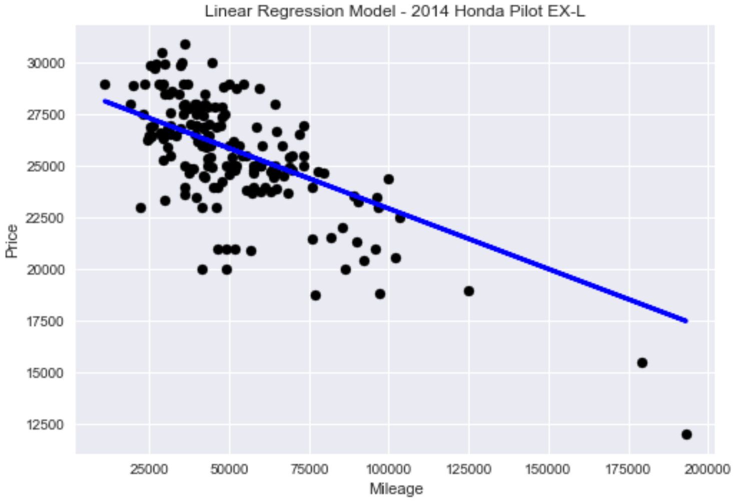
Figure 22: Honda Pilot LX Linear Regression

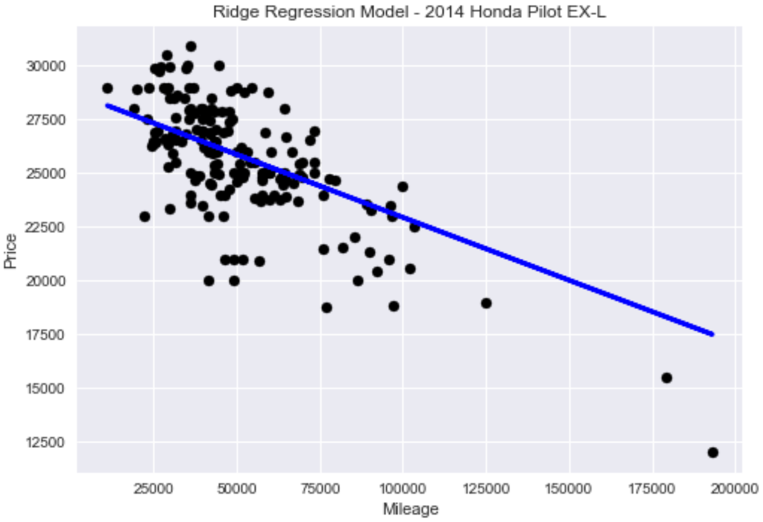


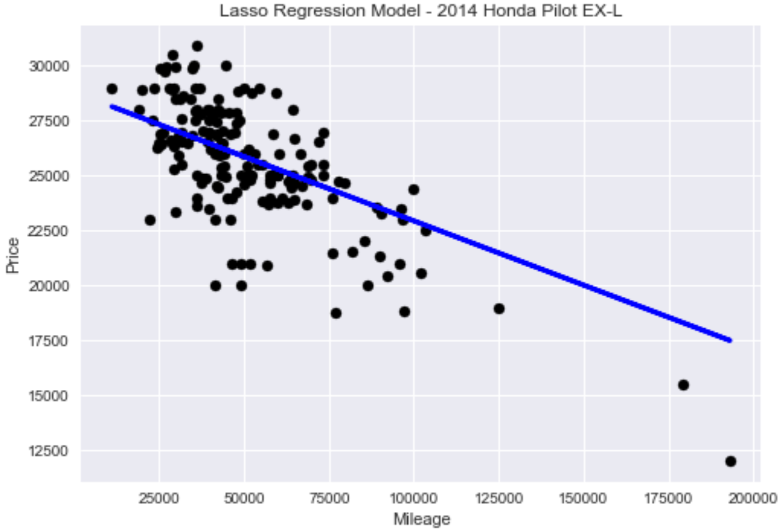


Regularization Techniques Utilized

Regularization techniques were employed using the Lasso Regression and Ridge Regression algorithms in Sci-kit Learn, but their results did not yield any change in the rMSE or regression line. This is due to the fact that the model we have is univariate: mileage is the only variable used to estimate price in our linear model. Other features such as year, make, and model were already controlled for and removed prior to creating the original linear regression. Perhaps the only additional data we could have used to create a multiple linear regression would be other metrics that were not captured in the dataset such as tire quality, mechanical condition, maintenance history, etc. Below are the plots generated from Honda Pilot EX-L data using linear regression, lasso regression, and ridge regression models in Sci-kit Learn:







Future Directions

Used vehicle valuations are a unique challenge in that the number of features necessary to build an accurate model is very high. Unfortunately with the datasets I had access to, I was only able to use Year, Make, Model, Trim, Price, and Mileage as features. Given a more detailed dataset that included values for features such as maintenance history, primary use, mechanical condition, and factory options, it would be possible to take advantage of other machine learning algorithms and tools to produce price estimates that are narrower and more realistic. This would allow Beepi to make more precise and informed business decisions about its revenue stream.

Final Thoughts and Caveats

After examining Beepi's dataset and the supplementary Kaggle/TrueCar dataset, I would recommend that the company make several key changes:

1. Revise the linear pricing models used to value vehicles that are costing the company money or halt transactions on these vehicles until it is possible to do this
2. Acquire as detailed of a dataset as possible that contains actual used vehicle transactions from the United States
3. Use customer purchasing habits to guide marketing campaigns and decisions

The primary challenge that I encountered when analyzing Beepi's dataset was its size; it was far too small to provide a basis for machine learning models. For instance, when I isolated vehicles based on their Year/Make/Model/Trim, only a small handful had more than 30; none had more than 40 data points. I supplemented the Beepi dataset with the Kaggle/TrueCar dataset, which contained listing prices for over 850,000 different vehicles in the United States. This provided the much needed data to build reasonable linear regression models.

The secondary challenge that I encountered was the univariate nature of the linear regression models that I created. After I isolated vehicles based on their Year/Make/Model/Trim, there were no features other than their mileage and price; the linear models created use only mileage to estimate price. Due to this, other machine-learning algorithms, such as Lasso Regression and Ridge Regression, yielded no improvements over the simple linear regression model used.

Finally, the major caveat that I must address is the nature of the data contained in the Kaggle/TrueCar dataset. It is comprised solely of *listing prices*, meaning that they are not final transaction prices. Although the principles used to create the linear regression models would be the same with actual transaction data and could be used if actual transaction data is fed into them, the client should not use the final results of this model generated with this data. The main reason is that used vehicles are often marked up by sellers in anticipation of price negotiations the buyer may wish enter in order to lower the purchase price of the vehicle. This likely means that the prices in the Kaggle/TrueCar dataset are inflated relative to their market value. Practically, this would mean if Beepi were to incorporate the models from this report into their pricing pipeline, it would likely overvalue sellers' vehicles and generate losses if buyers refuse to purchase cars at an inflated price from Beepi. With a dataset comprised of actual transactions, the models would yield useful information that Beepi could introduce into its customer-facing production pipeline.