REGRESSION ANALYSIS ON NEW VEHICLE SALE PRICE

William Roberts

1 BACKGROUND AND DESCRIPTION OF PROBLEM

Kelley Blue Book is a vehicle valuation and automotive research company based in Irvine, California. The company provides market value prices for new and used automobiles of all types. Kelley Blue Book products include Fair Purchase Price and Fair Market Range. These value types are based on actual transactions of what individuals are paying for vehicles and are adjusted accordingly as market conditions change on a weekly basis. For used cars, KBB also provides suggested retail value, trade-in value, private party value, certified pre-owned value, and auction value. As technology becomes more accessible and affordable, new players are popping up in the vehicle valuation industry. Companies like TrueCar and Carvana introduce their own values to compete in the already saturated industry. It is for this reason that Kelley Blue Book must stay competitive by exploring other methods that will produce more accurate and reliable values for the consumer.

Since Kelley Blue Book receives data from several source including dealers, auctions, and other external partners, it would be ideal to train a statistical model to predict the sale prices of vehicles. The model will utilize new car transaction data compiled on a weekly basis from an existing external partner. The output of this model will serve as the Fair Purchase Price for new vehicle purchases.

2 DATA AND DESCRIPTION OF FEATURES

New vehicle transaction data was gathered from one of Kelley Blue Book's external partners. The data consists of over 285,000 new cars and trucks sales from December 2016. While not all of the features were used to determine the model, it is important to understand how the data is structured. Let's take a look at some of these features:

Model Year:

The model year of the vehicle sold. Since the data was for transactions in December 2016, we expected most of the model years to be 2016s and 2017s, and indeed this was the case. Interestingly enough, there were also some model year 2014s and 2015s scattered throughout the data.

Make Name:

The manufacturer of the vehicle sold. Nissan leads all makes with over 32,000 new car transactions in December while Alfa Romeo had only one transaction. Exotic manufacturers such as Lamborghini and Ferrari were excluded in the data because it is unusual to see these vehicle sell for under MSRP. Therefore, it would not be considered Fair Market value.

Drivetrain:

The drivetrain configuration of the vehicle sold. These include four-wheel drive (4WD), all-wheel drive (AWD), front-wheel drive (FWD), Rear-wheel drive (RWD), and two-wheel drive (2WD). Depending on the vehicle, the drivetrain configuration might be the same but have different names. For example, 2WD can be either FWD or RWD. However, 4WD and AWD are different due to the way power is distributed to the wheels.

Mileage:

The mileage of the vehicle sold. Since these are new car transaction, we expect the mileage to be relatively low across the data. While the majority of the vehicles sold had less than 100 miles, there were vehicles sold with several hundred miles.

Days in Inventory:

The number of days the vehicle stood in the dealer's inventory before being purchased. This ranged from less than 1 day to several hundred days with an average of just over 76 days.

Number of transmission speeds:

The number of transmission speeds the vehicle had. Certain vehicles, such as hybrids, have a special type of transmission called Continuous Variable Transmission (CVT) in which there are no speeds. In this case, we simply set the number of speeds to one.

Transmission type:

The transmission of the vehicle sold, either manual or automatic. We encoded this feature to set automatic as one and manual as zero.

Door count:

The number of doors the vehicle had.

Engine Cylinders:

The number of engine cylinders the vehicle had. Hybrids and Electric vehicles again gave us trouble here. Since they have both an electric and gas, the data was null for these vehicles. In this case, we simply set the cylinders to zero.

Engine Displacement:

The swept volume of pistons inside the engine cylinders measure in liters. We encountered similar issues here like with Engine Cylinders.

Engine Type:

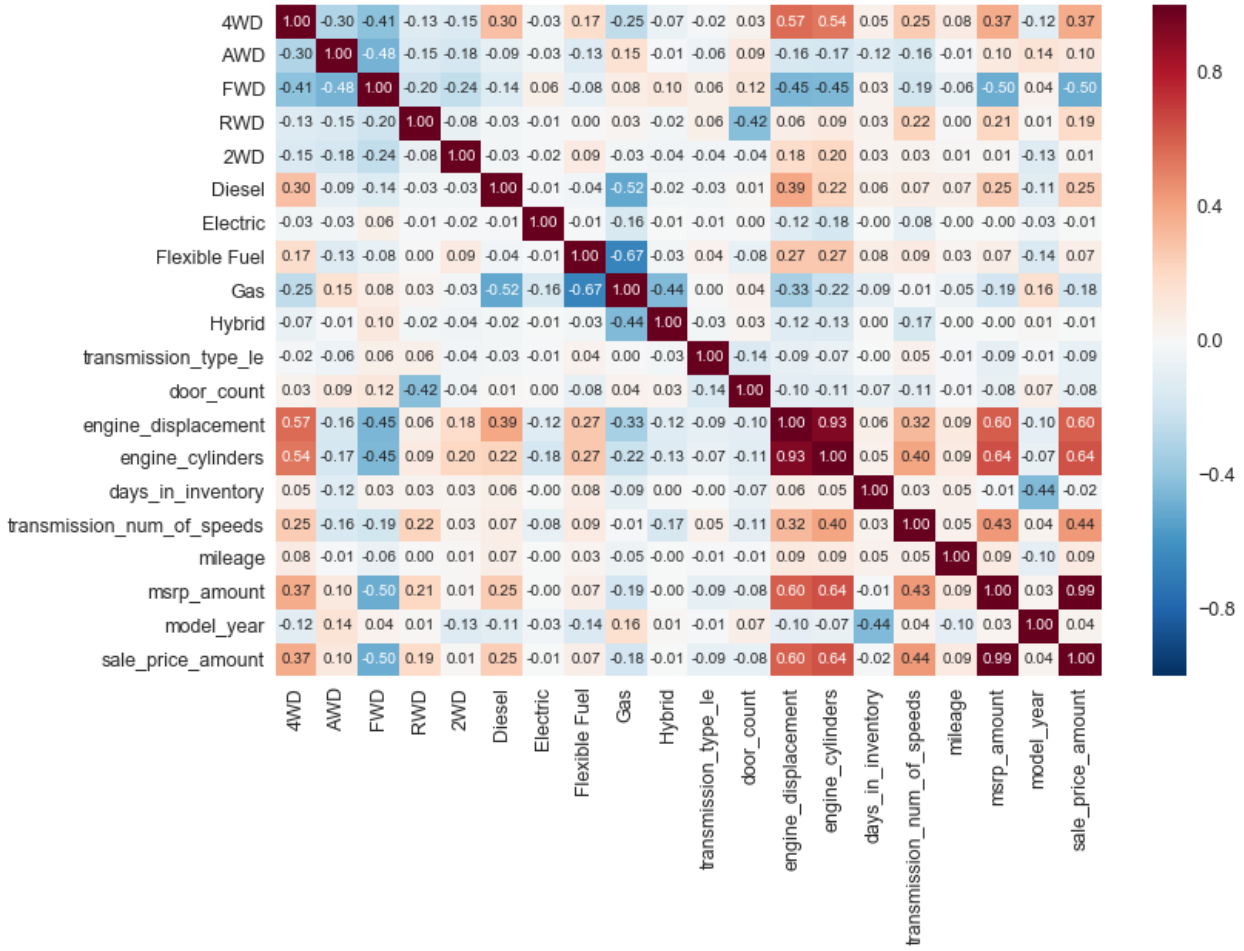
These include gas, diesel, hybrid, electric, and flexible fuel. We created dummy variables to set the various engine types here.

Sale Price Amount:

The amount the new car was sold for, also the target we are predicting.

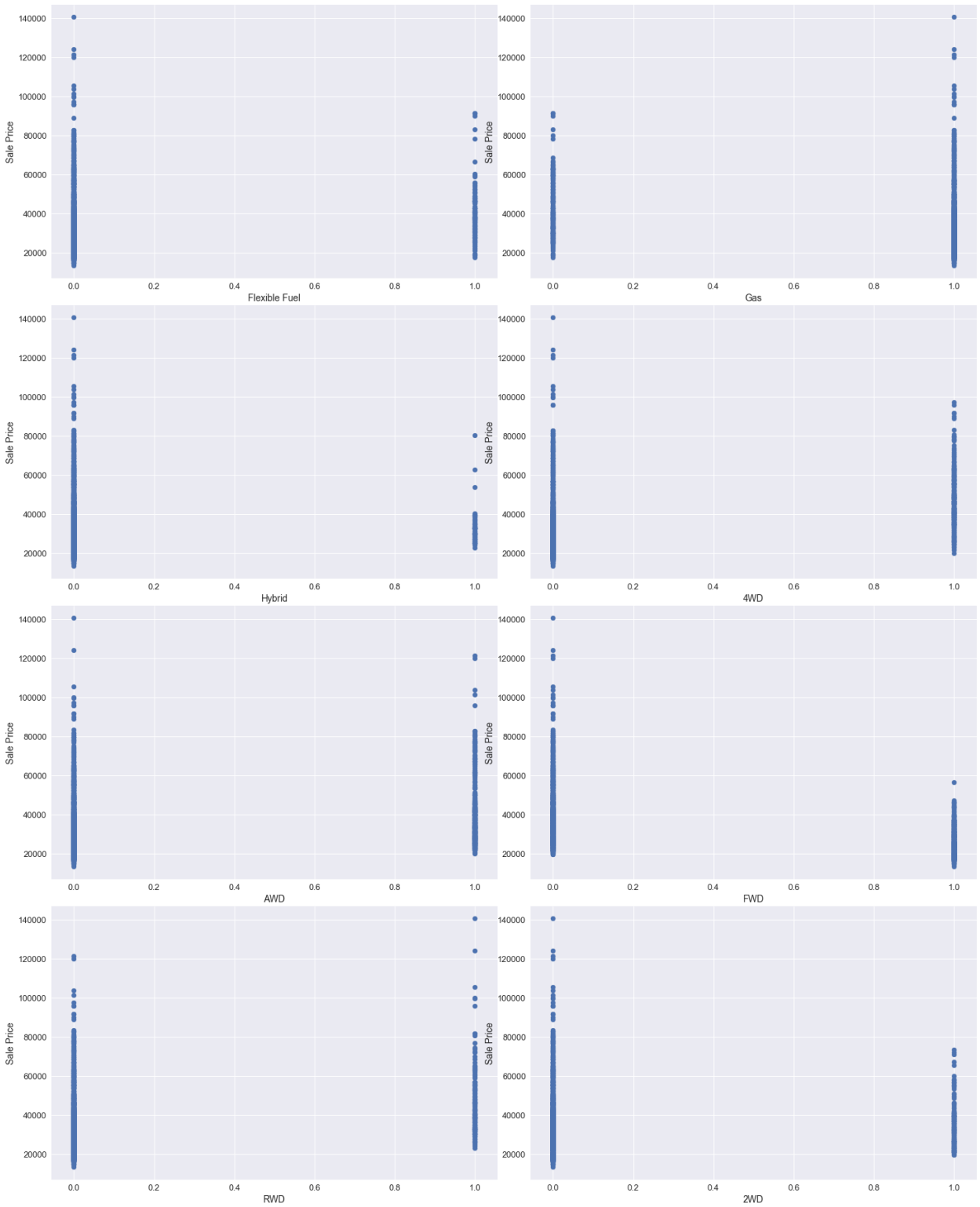
3 Feature Selection and Analysis

We will use a correlation matrix to quantify the linear relationship between each of the features and the target, the sale price amount of the vehicle. This also allows us to visualize any features that are highly correlated, also known as multicollinearity. While multicollinearity will not reduce the predictive power or reliability of the model as a whole, we might want to reduce it in order to improve the interpretability.



Looking at the correlation matrix, it appears that MSRP is highly correlated with the sale price of the vehicle. This makes sense since, usually, the higher the MSRP the high the sale price of the vehicle. We might want to consider dropping this feature. Other than FWD, there isn't any features that have strong negative linear correlations with sale price.

We can further examine the relationships of each of the features to the sale price by looking at the scatter plots of each feature.


Sale Price
Flexible Fuel
Gas
Hybrid
4WD
AWD
FWD
RWD
2WD

4 REGRESSION ANALYSIS

We first trained a linear regression model to establish a baseline of how well it predicted the target variable. However, after analyzing the results of the model, we discovered that the condition number is large which might indicate that there is strong multicollinearity. Therefore, we also trained a Ridge regression and a LASSO regression to see if it would predict the target variable with more accuracy.

To measure the accuracy of the models, we utilized the coefficient of determination and the mean squared error.

Coefficient of Determination (R-Squared):
The coefficient of determination indicates how much of the variance in the target variable is explained by the model. In other words, it measures the goodness-of-fit of the model. It is given as a percentage with a high value indicating that the model explains all the variability well while a low value indicating that the model doesn't explain the variability well.

However, R-squared does not indicate whether a regression model is adequate. For example, while a model with high R-squared might explain the variance in the target variable well, it might generalize well to future data due to overfitting. As a result, we will also be another metric to assess the accuracy of the model.

Mean Squared Error (MSE):
The mean squared error is a measure of the averages of the squares of the errors or deviations. In other words, the difference between the estimator and what is estimated. It is always non-negative, and values closer to zero are better. The MSE is useful for comparing different regression models or for tuning their parameters via grid search and cross-validation.

Let's examine the output of the linear regression model from statsmodels:

OLS Regression Results

Dep. Variable:	sale_price_amount	R-squared:	0.988
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	8.855e+05
Date:	Mon, 29 May 2017	Prob (F-statistic):	0.00
Time:	12:19:15	Log-Likelihood:	-1.7703e+06
No. Observations:	199810	AIC:	3.541e+06
Df Residuals:	199791	BIC:	3.541e+06
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
4WD	-4.277e+05	1.82e+04	-23.530	0.000	-4.63e+05	-3.92e+05
AWD	-4.276e+05	1.82e+04	-23.527	0.000	-4.63e+05	-3.92e+05
FWD	-4.28e+05	1.82e+04	-23.545	0.000	-4.64e+05	-3.92e+05
RWD	-4.288e+05	1.82e+04	-23.591	0.000	-4.64e+05	-3.93e+05
2WD	-4.278e+05	1.82e+04	-23.539	0.000	-4.63e+05	-3.92e+05
Diesel	-40.9374	1704.367	-0.024	0.981	-3381.456	3299.581
Electric	-881.9053	1706.037	-0.517	0.605	-4225.697	2461.887
Flexible Fuel	201.4585	1704.314	0.118	0.906	-3138.957	3541.874
Gas	47.8528	1704.252	0.028	0.978	-3292.441	3388.146
Hybrid	-158.0088	1704.483	-0.093	0.926	-3498.755	3182.737
transmission_type_le	185.5844	21.491	8.635	0.000	143.462	227.707
door_count	-357.3567	10.617	-33.660	0.000	-378.165	-336.548
engine_displacement	-152.9995	9.290	-16.470	0.000	-171.207	-134.792
engine_cylinders	50.2641	8.362	6.011	0.000	33.875	66.653
days_in_inventory	-1.4115	0.057	-24.706	0.000	-1.523	-1.299
transmission_num_of_speeds	51.4134	1.722	29.857	0.000	48.038	54.788
mileage	0.0297	0.041	0.728	0.467	-0.050	0.110
msrp_amount	0.9331	0.000	2558.195	0.000	0.932	0.934
model_year	212.8830	8.976	23.717	0.000	195.290	230.476

Omnibus:	22123.624	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	112616.937
Skew:	-0.426	Prob(JB):	0.00
Kurtosis:	6.578	Cond. No.	4.43e+08

The results shows a high R-squared of 0.988 which indicates a strong goodness-of-fit. This model also produces a MSE of 2,859,187.35. The MSE by itself doesn't tell how accurate the model is. We will need to calculate the MSE of each of the models to compare which is better.

Let's compare the R-squared and MSE of each of the models:

Model	R-squared	MSE
Linear	0.988	2,859,187.35
Ridge	0.988	2,859,186.69
LASSO	0.988	2,859,186.72

As we can see, all three models produces the same R-squared, however, the MSE for Ridge regression is the lowest.

We can also examine the significance of the coefficient to see if the performance of the model can be improved by dropping certain features. If the p-value is below a certain threshold (usually 0.05), then we can assume that the computed coefficient is statistically significant. Reviewing the output above, 6 features have p-values greater than 0.05 (Diesel, Electric, Flexible Fuel, Hybrid, Gas, and Mileage). Let's remove these features and see if it improves the model.

OLS Regression Results

Dep. Variable:	sale_price_amount	R-squared:	0.988
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	1.327e+06
Date:	Mon, 29 May 2017	Prob (F-statistic):	0.00
Time:	14:51:47	Log-Likelihood:	-1.7704e+06
No. Observations:	199810	AIC:	3.541e+06
Df Residuals:	199797	BIC:	3.541e+06
Df Model:	12		
Covariance Type:	nonrobust		

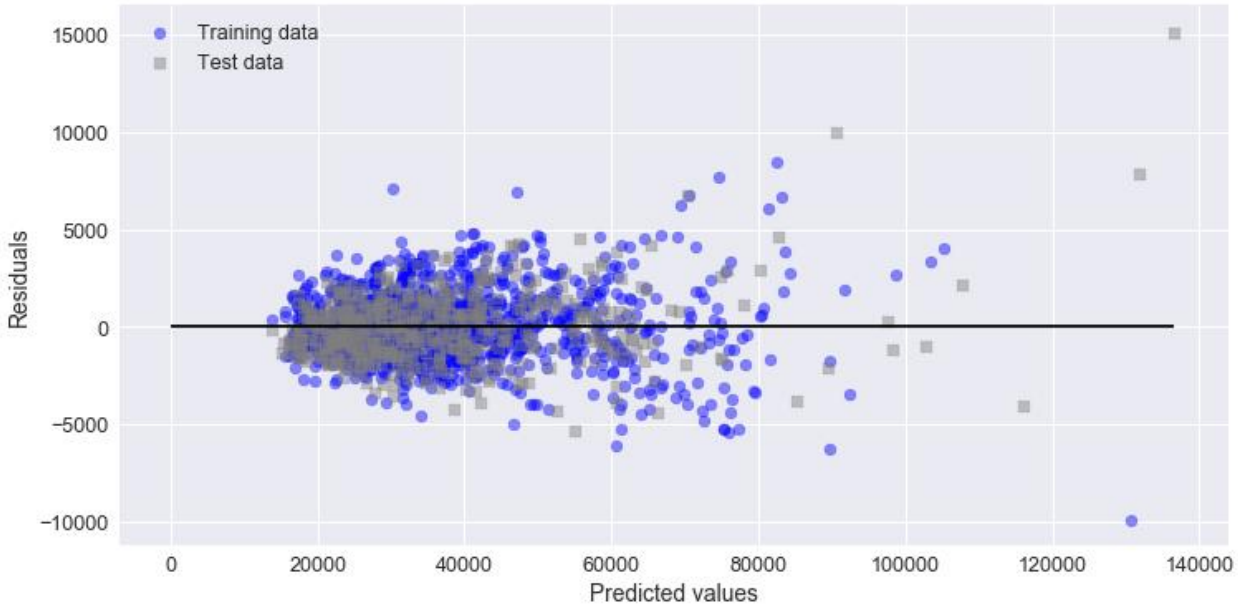
	coef	std err	t	P>|t|	[0.025	0.975]
4WD	-4.248e+05	1.79e+04	-23.734	0.000	-4.6e+05	-3.9e+05
AWD	-4.247e+05	1.79e+04	-23.729	0.000	-4.6e+05	-3.9e+05
FWD	-4.251e+05	1.79e+04	-23.749	0.000	-4.6e+05	-3.9e+05
RWD	-4.259e+05	1.79e+04	-23.795	0.000	-4.61e+05	-3.91e+05
2WD	-4.249e+05	1.79e+04	-23.742	0.000	-4.6e+05	-3.9e+05
transmission_type_le	195.9909	21.469	9.129	0.000	153.912	238.069
door_count	-359.4518	10.605	-33.893	0.000	-380.238	-338.665
engine_displacement	-171.9498	8.070	-21.306	0.000	-187.768	-156.132
engine_cylinders	88.3518	7.376	11.979	0.000	73.895	102.808
days_in_inventory	-1.4013	0.057	-24.519	0.000	-1.513	-1.289
transmission_num_of_speeds	54.7308	1.695	32.295	0.000	51.409	58.052
msrp_amount	0.9314	0.000	2689.850	0.000	0.931	0.932
model_year	211.4255	8.875	23.822	0.000	194.030	228.821

Omnibus:	22087.315	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	113731.298
Skew:	-0.421	Prob(JB):	0.00
Kurtosis:	6.599	Cond. No.	4.36e+08

It didn't seem like removing those features made a significant improvement to the model. The R-squared remain at 0.988. In addition, the MSE of the model worsened to 2,871,500.95.

Residual Plots:

Residuals, or errors in the regression, are expected to be random and not display any pattern. If they do, it could indicate that the model is failing to account for all the deterministic variance in the data. In general, the errors are supposed to capture the noise which should be random and stochastic. Also, the errors should be normally distributed. Looking at our residual plot, there doesn't appear to be any specific pattern and the distribution is more or less normal.



5 CONCLUSION

The analysis looked at improving the accuracy of the new car sale price Kelley Blue Book provides in order to build consumer trust and improve its reputation as an innovator in the industry. We achieve this by building a statistical model utilizing several features that are identified as being correlated with the sale price.

Three models were trained and its performance were compared using R-squared and Mean Squared Error. All three models achieved similar R-squared of 0.988, however, Ridge regression was able to achieve the best MSE of 2,859,186.69.

Additionally, the significance of the coefficient were examine to see if the performance of the models would improve by removing these features. We identified six features with p-values of greater than 0.05 and were therefore dropped. However, the performance of the model did not improve after dropping these feature, achieving a MSE of 2,871,500.95 and a similar R-squared.

6 FUTURE WORK

As a future scope, the analysis could include a seasonality component to further improve the model. Vehicle sales tend to increase during specific times of the year, such as during March and April when people receive their tax refunds or during the holidays when manufacturers have aggressive sales events. This will no doubt impact the sale price of a vehicle. Additionally, the analysis can be further expanded to include used vehicle sale price where there is even more ambiguity in vehicle values. Achieving success in this area will lead to even more consumer trust and domination of the market.