Identifying Amazon Influencers and The Effect They Have

Team 4: Ke Zang; Di Yao;
Shimiao Li; Luke Towers; Yue Wu

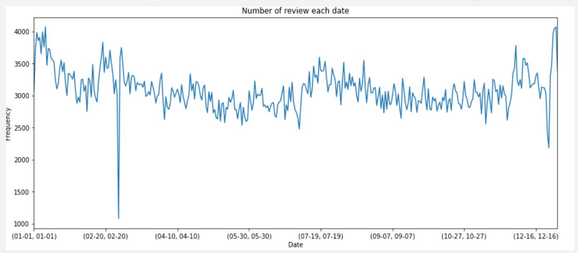
Description: Our capstone project aims to *identify amazon influencers* from a dataset of 250 million reviews and product descriptions. We utilize machine learning to predict influencers based on several variables including the sentiment of review text and length of reviews along with 5 other variables and estimate the effect on sales

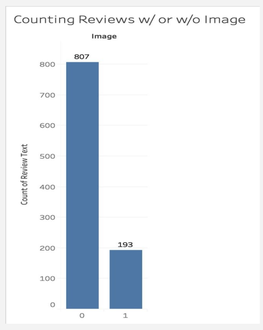
Data Preparation

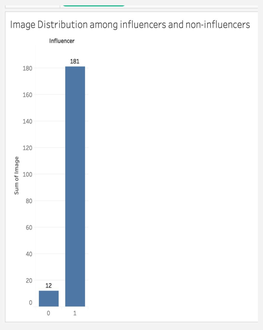
- **Raw Dataset:**
- This **"Cell Phone and Accessories"** review dataset has 1,128,437 rows and 12 columns. This category contains 48,186 products.
- **Data Cleaning:**
- Stropped out N/As and unused columns
- Convert data type to necessary data format
- **New Column Influencer:**
- Descended each review's annual helpful votes and selected top 500 customers as "influencers"
- Randomized 500 other users as "non-influencers

Exploratory Dataset Analysis

- Overall clients tended to leave higher average rating, influencers have a lower average rating than overall.
- In year 2011 and 2012, customers leave comments with highest number of words while influencers left highest at year 2010 and 2012.
- Almost 70% of the ratings are 5.0, which means customers are generally satisfied with the product.



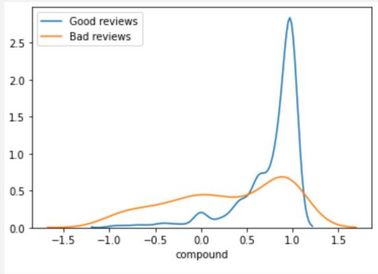




Method & Result

1)Sentiment Analysis

- Even if people give low ratings (1-3), their emotional can still be positive. The negative emotion of low rating (bad reviews) is higher than high rating (good reviews).
- Highest positive and negative sentiment reviews. From the table, we can see the positive and negative emotion text with highest score are from non-influencer.
- Therefore, we think people like someone's review are not influenced by the emotion of the commentator.

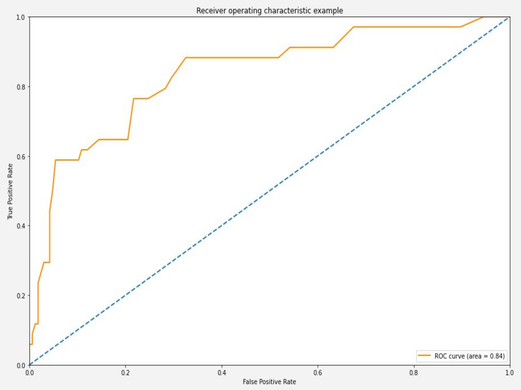


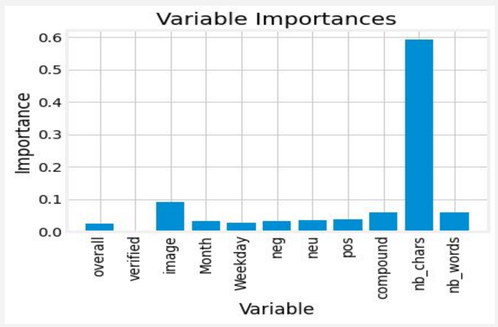
	review_clean	pos	influencer
823	nicely build elegant good fit	0.919	0
572	fit perfectly protect well look good	0.852	0
893	good product good price good service	0.744	0
876	excellent packaging awesome look	0.722	0
917	great product nice price	0.697	0
739	beautiful product compact efficient functional	0.691	0
880	cool work great	0.690	0
743	love fun phone super durable	0.661	0
699	great prop easy put away	0.646	0
796	good price quick ship work great	0.636	0

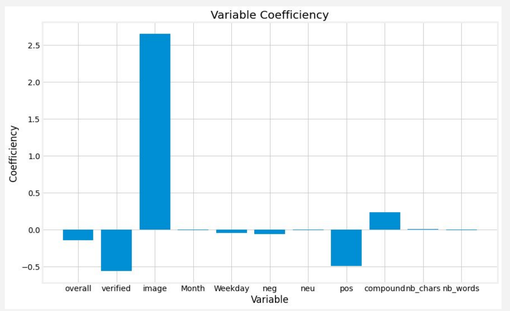
	review_clean	neg	influencer
929	suck waste money	0.655	0
517	screen go bad within day	0.467	0
978	absolutly horrible week use stop work stop rea...	0.456	0
524	super product problem	0.415	0
672	holster clip break real quick	0.412	0
706	week start fall apart quality suck	0.386	0
904	break first time drop phone	0.359	0
642	long hands-free	0.355	0
790	get disgust even return refund phone still bra...	0.329	0
841	place lexus yet	0.306	0

2)Machine Learning models in predicting influencers

- Random Forest: accuracy 57% , best indicators are number of words and image
- Logistic Regression: accuracy 87.6%, best indicator is image
- More sophisticated linear regression neural network model with 72% accuracy

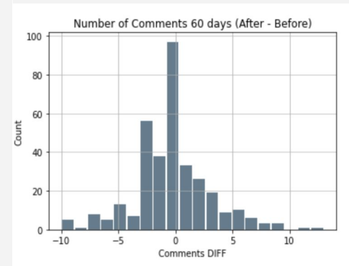


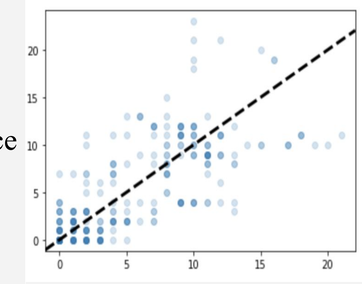




Impact on Sales

- **Assumption 1:** Sales is positively correlated to the number of comments for each product.
- **Assumption 2:** Compare the number of comments 60 days before and after
- **Conclusion:** The median of difference between 60 days before and 60 days after is 0. Probably not huge impact on sales.





Conclusion & Result

- The most accurate ML model is the logistic regression model. The best drives to indicate whether someone is an influencer or not are image, verified or not, compounded sentiment score and positive emotion.
- "influencers" reviews more based on whether they have an image or not and number of words instead of their positive or negative emotion.
- It is clear that without effective sales data of amazon products it is difficult for us to determine whether these "influencers" really affect the sale of products.

Reference: https://towardsdatascience.com/detecting-bad-customer-reviews-with-nlp-d8b36134dc7e https://github.com/jeremygrace/amazon-reviews