FASHION TREND FOR INNERWEAR

Drexel University
Group 2: Yue ‘Alex‘ Fu, Jiameizi Yao, Camila Pareja, Tianyue 'Florence' Wang
Winter 2020

Agenda

Fashion Brands

Business Problems

Our Dataset

Product Classifications

Prediction Models

Performance & Conclusion

Limitations

FASHION BRANDS

- Wacoal
- B.tempt'd
- Victoria's Secret
- Calvin Klein
- Hanky Panky

Brands

Wacoal

- Founded in Japan in 1949 by Koichi Tsukamoto.
- Goal = Beautiful intimate apparel + high quality + innovation.
- Wacoal launched the brand in the US in 1985 that showcased high quality intimate apparel that was comfortable and fit different types of bodies.
- A company that supports breast cancer awareness worldwide, they launched in 2001 "Fit for cure" an initiative offering complementary bra fittings at more than 1000 stores in the US and Canada. This initiative's goal is to raise money to fight breast cancer. Today, the company has donated more than $4.7 million dollars.
- This brand is committed to helping women find the perfect bras, those that fit and feel comfortable. Therefore, they have trained a group of consultants in order to help women find the perfect bra. These consultants are available at all times in their website.
- Because this brand focuses on all body types, they have a wide range of styles and sizes, ranging from A – I cup. Without compromising quality and comfort.

References: https://www.wacoal-america.com

B.tempt'd

- This brand is also part of Wacoal family. However, they have a smaller range of sizes than Wacoal.
- Every bra and panties are fitted on real women; Hence the fit is the most important goal of this brand.
- Sizes range from A-DDD.
- B.tempt'd focuses on empowering women through modern, comfortable, and creative pieces to make women feel beautiful.
- They have been referred to as the "Best Selling Bra Brand in US Department Stores" since 2005.
- All exclusive collections are designed in NYC with a group of 38 women constantly working to create innovative and interesting intimates.

References: https://btemptd.wacoal-america.com/

Victoria's Secret

- Founded in Palo Alto, California in 1977 by Roy and Gaye Raymond.
- A company that sells underwear, women's clothing, lingerie, swimwear, footwear, fragrances, beauty products and make up.
- Les Wexner purchased VS in 1982 and expanded into shopping malls across US and other countries.
- The company has been widely known for their marketing strategies and annual fashion show featuring VS "Angels" that run up until 2018.
- In 2012 VS launched the "VS Designer Collection" described by Vogue as their first high end lingerie line.
- VS owns 1017 stores and 18 independently owned stores.
- Products are manufactured in India, Sri Lanka, Jordan and the Pacific island of Saipan.

References**:** https://en.wikipedia.org/wiki/Victoria's_Secret

Calvin Klein

- American company founded in 1968 by Calvin Klein and Barry K. Schwartz in NYC.
- This company specializes in leather, lifestyle accessories, home furnishings, perfumery, jewelry, underwear and ready-to-wear.
- Although the company almost went bankruptcy in 1992, CK managed to overcome this bump in the road with the success of highly popular underwear and fragrance lines.
- During 1990-1995 former CK designer John Varvatos pioneer a men's underwear called ”Boxer brief” a hybrid of boxer shorts and briefs that were made famous by being featured on Mark Wahlberg.
- CK was acquired by PVH Corp. in 2003.
- Signature CK Underwear boutiques can be found in Ljubljana, Buenos Aires, Mexico City, among other international locations.
- CK bra sizes range from A-DD and their panties from XS-3XL.
- CK underwear’s goal is to provide clients with comfortable and sporty designs made from a blend of cotton and spandex.

References: https://www.calvinklein.us/en & https://en.wikipedia.org/wiki/Calvin_Klein_(company)

Hanky Panky

- American brand founded in 1977 by designer Gale Epstein.
- Brand's name came from the idea of handmade lingerie crafted out of embroidered handkerchiefs.
- HP mainly focus is in underwear but also has a sleepwear line.
- The brand's goal is to provide products with a good fit, design, quality, comfort, and most importantly made in the US.
- In 1986 HP set a new standard for thongs and in 2004 the Wall Street Journal described their thongs as "Lace Butter" which created an immediate success among fashion magazines and celebrities.
- They are known for using top quality materials and responsible labor in order to create durable and high-quality products.
- HP signature lace styles are knitted in the US, with Supima Cotton fibers use to make crotch linings. All materials sourced in the US.
- Bra sizes range from XS-3X and panties from XS-XL.

Reference: https://www.hankypanky.com

BUSINESS PROBLEMS

Business Problems

What features should a product have in order to get good ratings?

Do women value more the quality of a product rather than its aesthetic when purchasing underwear ?

OUR DATASET

Data Characteristics

- **Describing Dataset**
  - Where has the data been gathered?
    - Kaggle.com (https://www.kaggle.com/PromptCloudHQ/innerwear-data-from-victorias-secret-and-others).
    - Products which are selling on Amazon.com.
  - What are the input and output variables?
    - Input variables: price, brand name, review count, sizes, etc.
    - Output variable: rating/rating level.
- **Descriptive Analysis**
  - Number of columns and rows
    - 14 Columns and 31612 observations.
  - Percent of missing values by columns mode
    - No missing values.

Data Cleaning/Preparation

- **Data Preprocessing**
  - Brand name:
    - Replace messy brand name using the first word in product name.
    - Simplify Calvin-Klein and B-tempt'd.
  - Color: Some missing available sizes are shown in the color, we move them to the available sizes.
  - Price: Take out $ sign and change them into numerical values.
  - Sizes: Create new variables that count the number of total sizes and available sizes, and the difference between them.
  - Creating new variables
    - Create dummy variables for the product category and the top key words in the description and style-attributes variables.
  - Dropping unnecessary variables
    - Drop the variables with zero and high cardinality.
    - Drop URL, product name, retailer, and the original variables which has been transferred to new ones.
  - Imputing missing values
    - Impute the missing values in color with mode due to the problem in available sizes above.

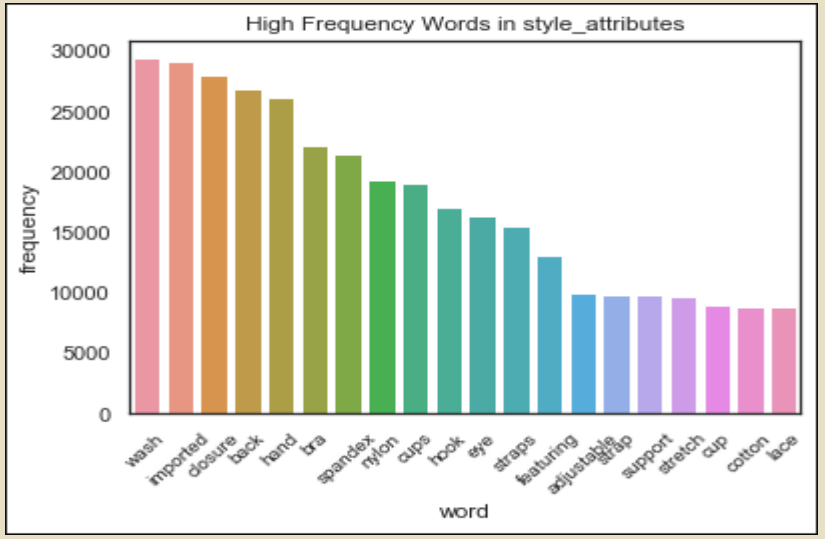
Mean, Medium and Mode

Variables	Mean	Medium	Mode
MRP	48.42	50	65
Price	44.28	46	65
Rating	4.26	4.3	4.4
Review count	464.49	380	377
Total sizes	26.09	25	23

Data Visualization (1)

- Word cloud for Description

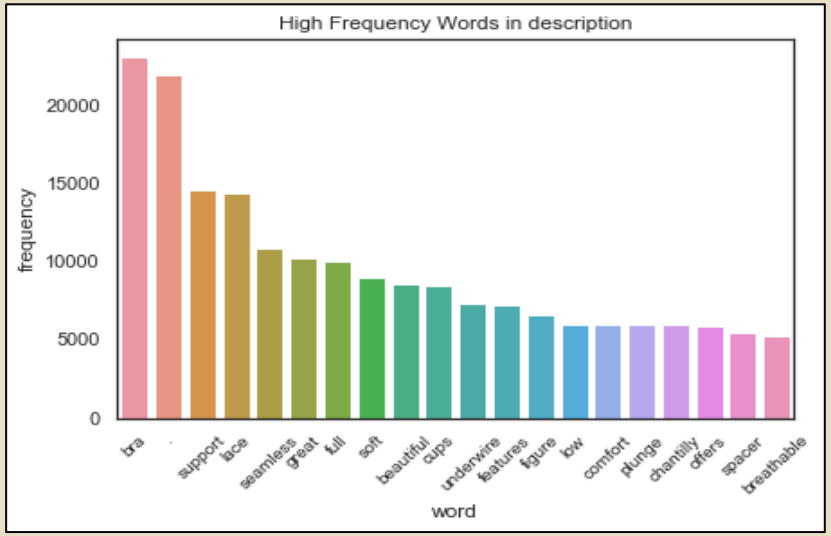
- Frequency for Description



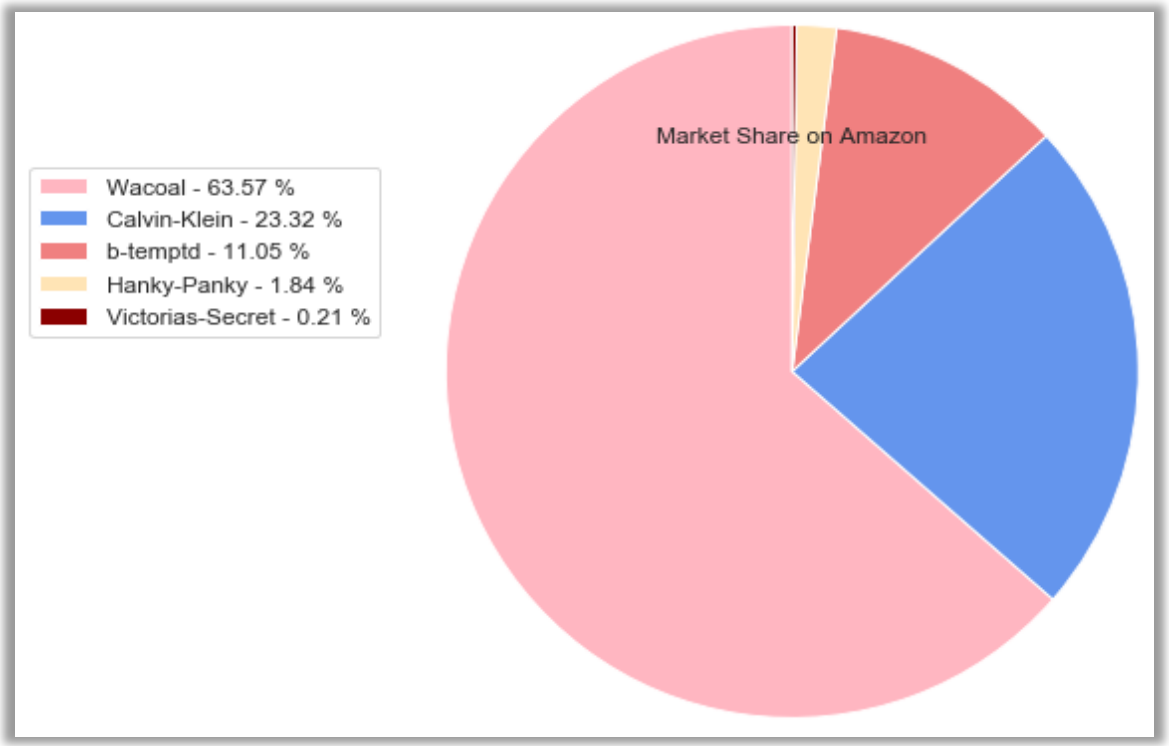
Data Visualization (2)

- Word cloud for Style Attributes

- Frequency for Style Attributes

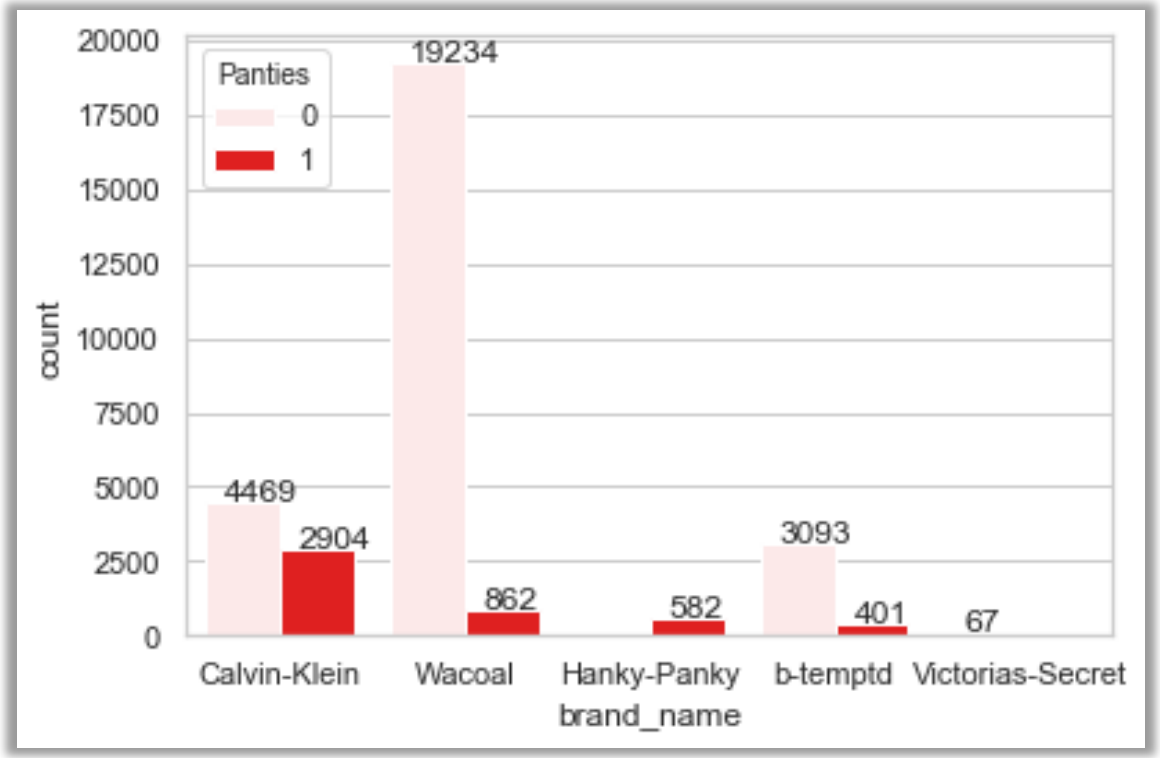


Market Share on Amazon



- As seen on the pie chart, we can see that the dominant brand in our dataset is ***Wacoal*** covering 63.57% of our entire data with information about both bras and panty sales on Amazon.
- ***Hanky Panky*** covers only 1.84% of our data with information only about panties.
- ***Victoria's Secret*** covers only 0.21% of our data with information only about bra sales.

Product Type for Brands

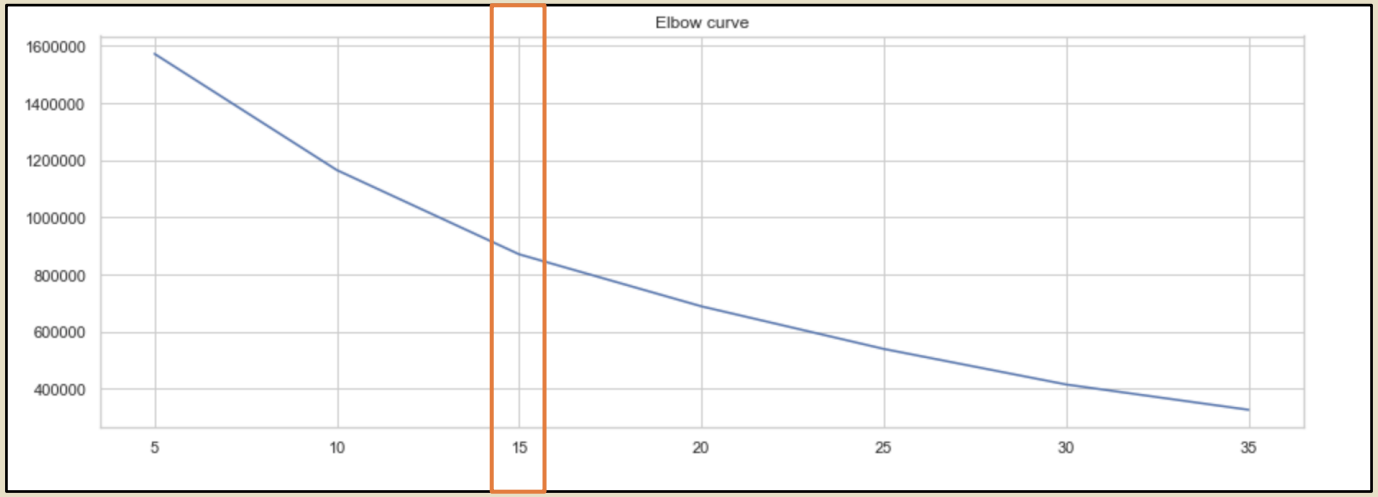


According to the chart on the left, we clearly see that not all companies have observations about "panties" and that most of our dataset focuses on bras.

Calvin Klein → 2,904
Wacoal → 852
Hanky-Panky → 582
B.tempt'd → 401
Victoria's Secret → 0

PRODUCT CLASSIFICATIONS

K Means Clustering



Elbow curve appears that 15 is the optimal number of classes.

K Means Clustering

- ***Cluster 1*** High in price but relatively low in mrp, low review_count. Prefer superior lift, support, contour, soft and comfortable features in description, T-shirt and sheer in style_attribute. Avoid breathable, straps, molded in description, support in style_attribute.
- ***Cluster 2*** Low in mrp, low review_count, small amount of Wacoal products. Prefer sling, fabrication, strapless in description. Low frequecy in breathable, soft, contour and molded in description, strap, smooth, strenth, support in style attribute.
- ***Cluster 3*** High in mrp but low in price. High review_count but low rating. High in total_sizes_num, available_size_num. Large number of Victoria's Secret, small amount of Calvin-Klein. High frequency in underwire, trim, stretch, support in style_attributes, comfort, chantilly, breathable, molded in description. Low frequency in sling, lightweight, modern, comfortable in description, straps, lace, sheer, elastane, hook, cotton, nylon in style_attributes.
- ***Cluster 4*** Low in mrp. Large number of Victoria's Secret. Prefer basic and soft in description, hook and scalloped in style_attributes. Avoid sling, breatheable, contour, molded in description, sheer, smooth in style_attributes.
- ***Cluster 5*** High in both price and mrp. High review_count and high in panties. Large amount of Wacoal and Hanky-Panky. Low in both diff_size_num and available_size_num. High frequency in sling, modern, breathable, comfortable, straps and molded in description, strap, sheer, elastane, bra, adjustable in style_attributes. Low frequency in seamless, stretch, spacer, soft and chantilly in description, stretch, trim, sling in style_attributes

K Means Clustering

- ***Cluster 6*** High in both price and mrp. High in review_count. Large amount of Wacoal products. Small amounts of panties. Low in diff_size_num and available_size_num. High frequency in strap, sheer, support in style_attributes, seamless, spacer, breathable, soft in description. Low frequency in chantilly, fabrication in description, trim, sling, t-shirt, spandex in style_attributes.
- ***Cluster 7*** High in mrp but low in price. High in rating. Large amount from Wacoal, a few Victoria's Secret and Hanky-Panky products. High in available_size_num but low in diff_size_num. High frequency in coverage, lace, basic, soft, strap, cotton, and t-shirt. Low frequency in seamless , stretch, modern, comfort, chantilly, sheer.
- ***Cluster 8*** High in mrp but low in price. High review_count. High in diff_size_num but low in available_size_num. Large amount of Wacoal. High in strap, seamless, stretch, spacer, underwire, sling, nylon, soft, contour, support, molded. Low in sheer, breathable, t-shirt, smooth.
- ***Cluster 9*** Low in both price and mrp. High in diff_size_num but low in available_size_num. Large amount of Wacoal and Hanky-Panky. High frequency in strap, lightweight, comfort, breathable, trim, nylon. Low frequency in sheer, support, soft and chantilly.
- ***Cluster 10*** Low in mrp. Large amount of b-temptd but a few Wacoal. Prefer cotton, sling, t-shirt, comfortable, adjustable. Less strap, seamless, chantilly, breathable and support.

K Means Clustering

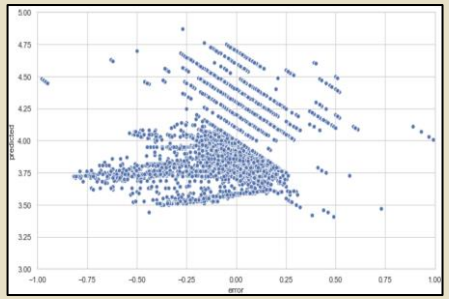
- ***Cluster 11*** Low in diff_size_num and available_size_num. Large amount of Calvin-Klein products but small amount of Wacoal and Hanky-Panky. High frenqency in sheer, strench, sling, smooth, soft, molded and spandex. Low frequency in seamless, modern, breathable, plunge, support, cotton, nylon and fabrication.
- ***Cluster 12*** Low in price. High in review_count. High in diff_size_num, available_size_num and total_sizes_num. Less pantities. Large amount of Victoria's Secret and Calvin-Klein but small amount of Hanky-Panky. More features related to lightweight, stretch, underwire, spacer, scalloped, chantilly, breathable, trim, spandex and adjustable. Less features related to seamless, sheer, comfortable and support.
- ***Cluster 13*** Low in both mrp and price. More features related to seamless, center, chantilly and straps. Less features related to sheer, breathable, strech, soft, support and molded.
- ***Cluster 14*** High in rating but low in review_count. Low in both mrp and price. Small amount of Wacoal. More features related to sling, hook, scalloped, chantilly, plunge, fabrication and molded. Less features related to strap, sheer, breathable, strtch, soft, smooth and support.
- ***Cluster 15*** High in mrp but low in price. High in review_count. Low in diff_size_num and available_size_num. Large amount of Wacoal. More features related to seamless, stretch, lace, comfort, center, breathable, trim, sling, soft, support and molded. Less features related to sheer, chantilly, contour.

PREDICTION MODELS

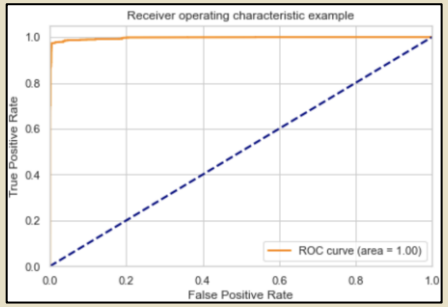
Five Models

- Before build the model, we transform the rating into binary as rating level, 1 means high rating for rating over 4.0, 0 for below 4.0. The rating level is the dependent variable for all models except liner regression.
- Besides, we found that the dataset is not balance. Therefore, we used SMOTE to oversample the dataset, making the number of both levels of the rating equal to 26793.

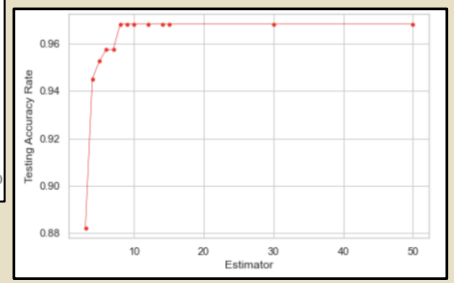
Linear Regression



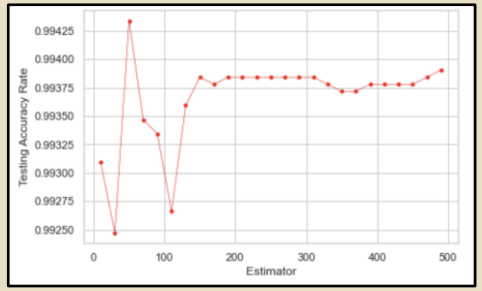
Logistic Regression



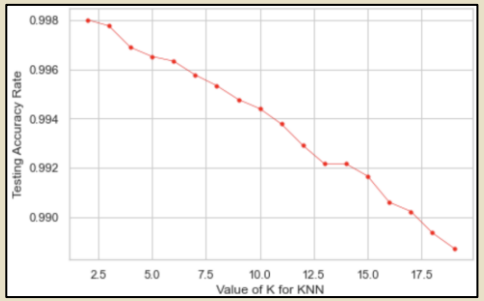
Decision Tree



Random Forest



K-NN



Linear Regression Model

description_breathable	description_comfortable	description_strapless	Panties	style_attributes_adjusta	description_chantilly	description_modern	description_molded
-0.10	0.12	-0.15	0.24	-0.05	-0.08	0.19	0.03
style_attributes_sheer	style_attributes_inner	description_contour	style_attributes_body	description_soft	description_basic	mrp	style_attributes_lace
0.06	0.03	0.00	-0.12	-0.03	0.11	-0.01	0.04
description_low	style_attributes_hook	style_attributes_elastan	style_attributes_nylon	description_coverage	style_attributes_cups	description_support	description_seamless
-0.11	-0.08	-0.09	0.18	0.16	-0.29	0.03	0.17
description_lightweight	description_spacer	style_attributes_center	diff_size_num	description_comfort	description_smooth	style_attributes_spande	total_sizes_num
0.04	0.07	0.04	0.01	0.20	0.01	-0.12	0.00
style_attributes_100	description_stretch	description_offers	description_full	style_attributes_shape	description_fabrication	description_great	style_attributes_underwire
0.02	0.24	-0.42	-0.01	0.10	0.46	0.01	-0.05
style_attributes_featurir	description_superior	Hanky-Panky	description_features	description_plunge	style_attributes_straps	style_attributes_scallop	description_lace
0.09	-0.08	-0.18	0.04	-0.08	0.14	-0.04	0.20
style_attributes_support	description_underwire	style_attributes_smooth	Victorias-Secret	b-temptd	style_attributes_cotton	style_attributes_t-shirt	description_sling
-0.08	0.06	0.02	0.83	0.01	0.09	-0.09	-0.26
Calvin-Klein	style_attributes_trim	description_figure	Wacoal	available_size_num	description_beauty	description_beautiful	style_attributes_back
-0.03	-0.10	-0.03	0.21	0.01	-0.38	-0.09	-0.03
style_attributes_stretch	description_straps	review_count	price	style_attributes_bra	description_.	style_attributes_strap	description_offer
-0.05	-0.18	0.00	0.00	0.05	-0.04	-0.16	0.17
style_attributes_sling	style_attributes_cup						
0.00	0.17						

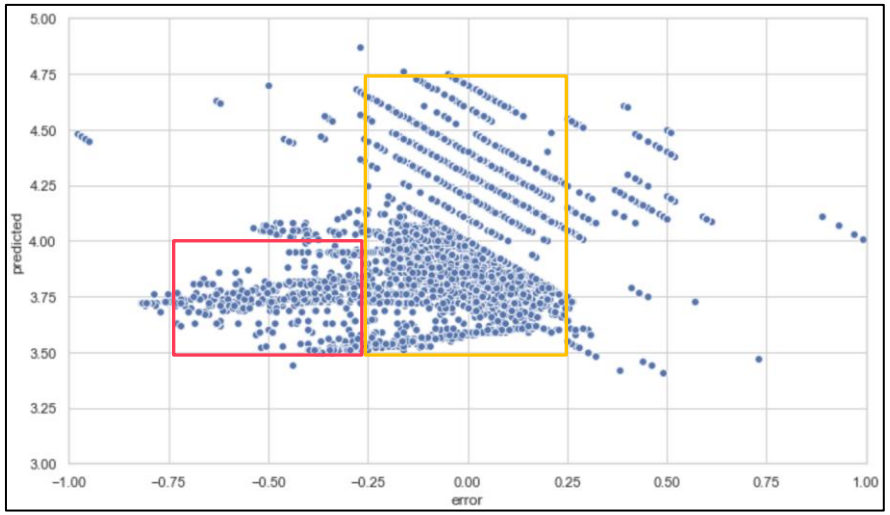
We built a linear regression model with R^2 equals to 82.76.

Coefficients:

- In the figure above, the green area means that the coefficients are positive and have a positive effect to rating, while the red area means the coefficients have a negative effect to rating.
- For example, Victoria's-Secret will get higher ratings on amazon. Customers are more satisfied by panties rather than bras. When there is 'beauty' in the product descriptions, the ratings are more likely to be low. We guess that it is because the brand focuses more on the appearance of the product but not the comfort. Additionally, price and review counts have no effect on ratings, which is out of our expectations.

Predicted Value vs Error Plot

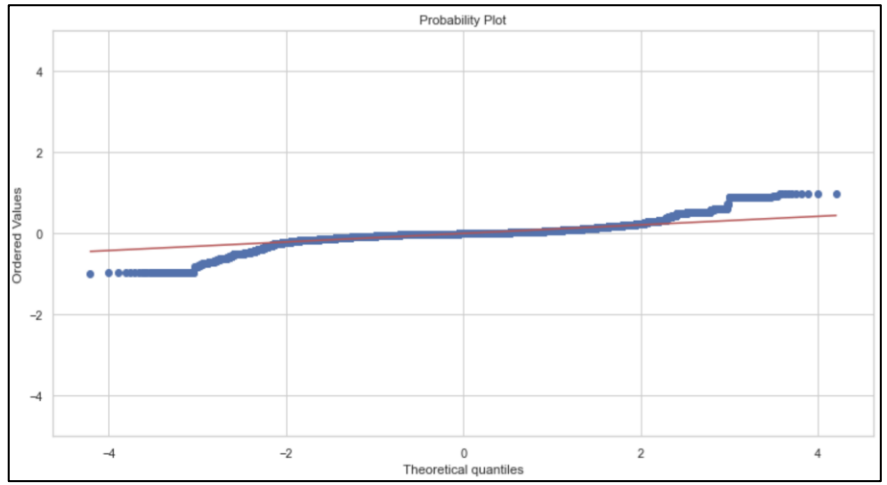
First plot is showing the predicted result and the error comparing with the actual value. We can see the most points are around 0 located between -0.25 and 0.25, which is highlighted by the yellow rectangle. When the predicted values are in the range of 3.5 to 4.25, it's more likely to have high negative error as shown in the red rectangle.



	actual	predicted	error
0	4.5	4.48	0.02
1	4.4	4.31	0.09
2	4.3	4.22	0.08
3	4.4	4.40	0.00
4	4.4	4.31	0.09
5	4.2	4.21	-0.01
6	4.2	4.04	0.16
7	4.7	4.25	0.45
8	4.7	4.20	0.50
9	4.2	4.21	-0.01

Probability Plot

From the Q-Q plot, we can see that our sample is normally distributed since it's fairly straight.



Reference: https://en.wikipedia.org/wiki/Normal_probability_plot

Variable Combinations Selection

- Comparing the two models in logistic regression, we find the second variable combination is better since the first one might be overfitting.
- Therefore, we apply other three models on both two variable combinations. Then we choose the better variable combination in the final modelling process and get the results.

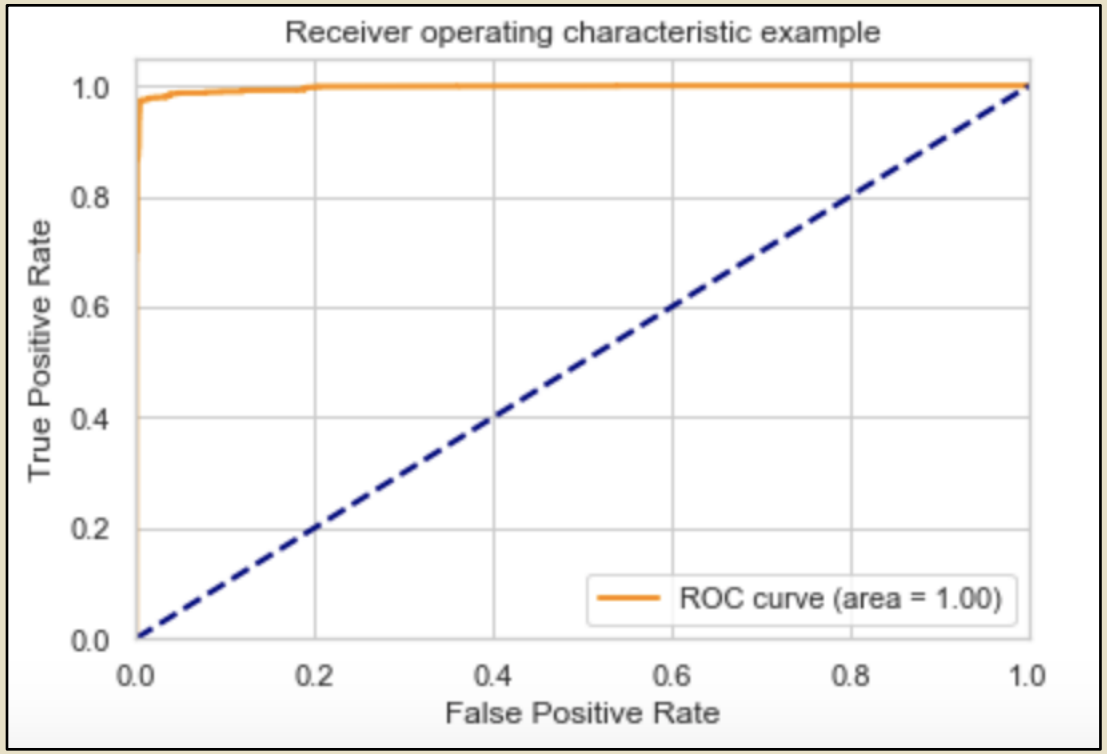
```
['description_soft',
 'description_coverage',
 'description_comfort',
 'description_full',
 'description_fabrication',
 'description_underwire',
 'Wacoal',
 'style_attributes_back',
 'style_attributes_bra',
 'style_attributes_strap']
```

VS

All 74 variables

- Logistic Regression
- Decision Trees
- Random Forest
- K-Nearest Neighbors

Logistic Regression Model (1)



	accuracy	sensitivity	specificity
Logistic Regression	0.978726	0.97962	0.97783

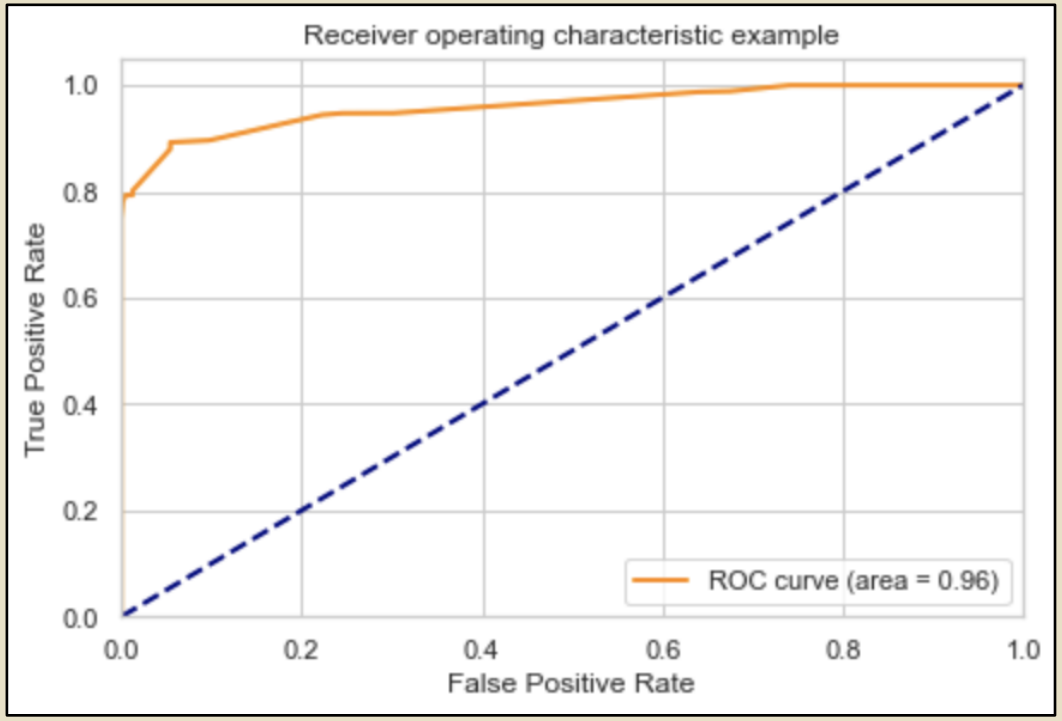
- Building the model by all 74 variables, we get the model with 97.9% accuracy.
- By applying the 5-fold cross validation, we get a more precise accuracy as 99.5%.
- ROC means the model is 100% better than randomly predicted results.

Logistic Regression Model (2)

After Feature Selection

We selected 10 variables as main features showing in the right screenshot.

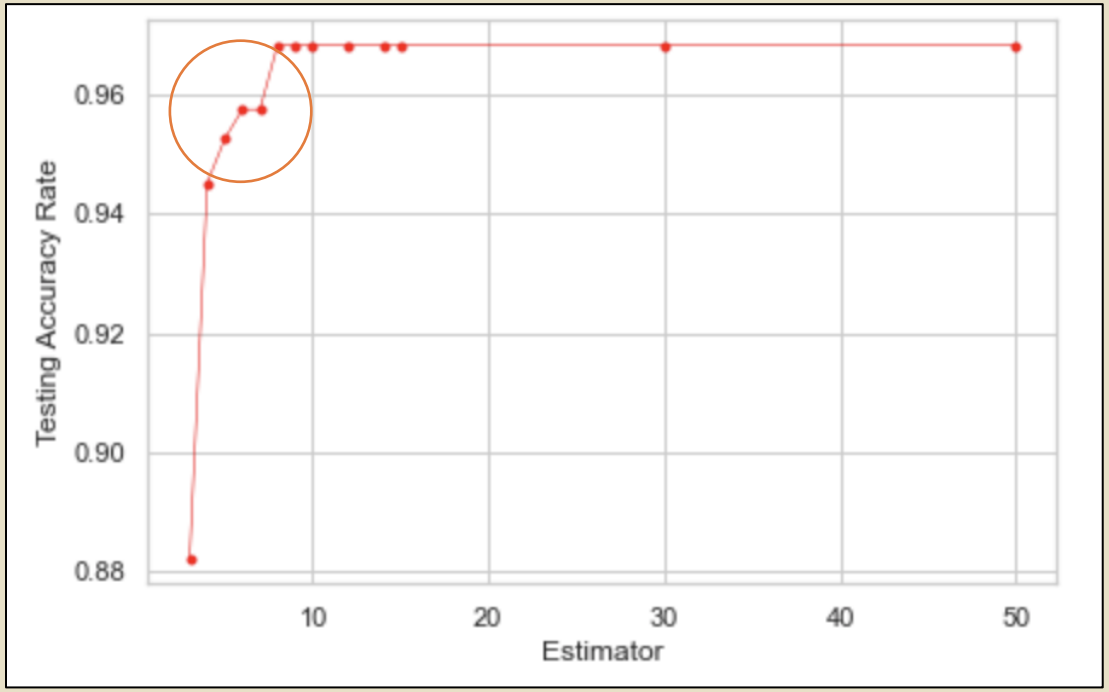
```
['description_soft',
 'description_coverage',
 'description_comfort',
 'description_full',
 'description_fabrication',
 'description_underwire',
 'Wacoal',
 'style_attributes_back',
 'style_attributes_bra',
 'style_attributes_strap']
```



	accuracy	sensitivity	specificity
Logistic Regression 2	0.918947	0.945321	0.892515

- Building the model with all 10 variables, we get the model with 91.9% accuracy.
- By applying the 5-fold cross validation, we get a more precise accuracy of 92.1%.
- ROC means the model is 96% better than randomly predicted results.

Decision Tree Model (1)

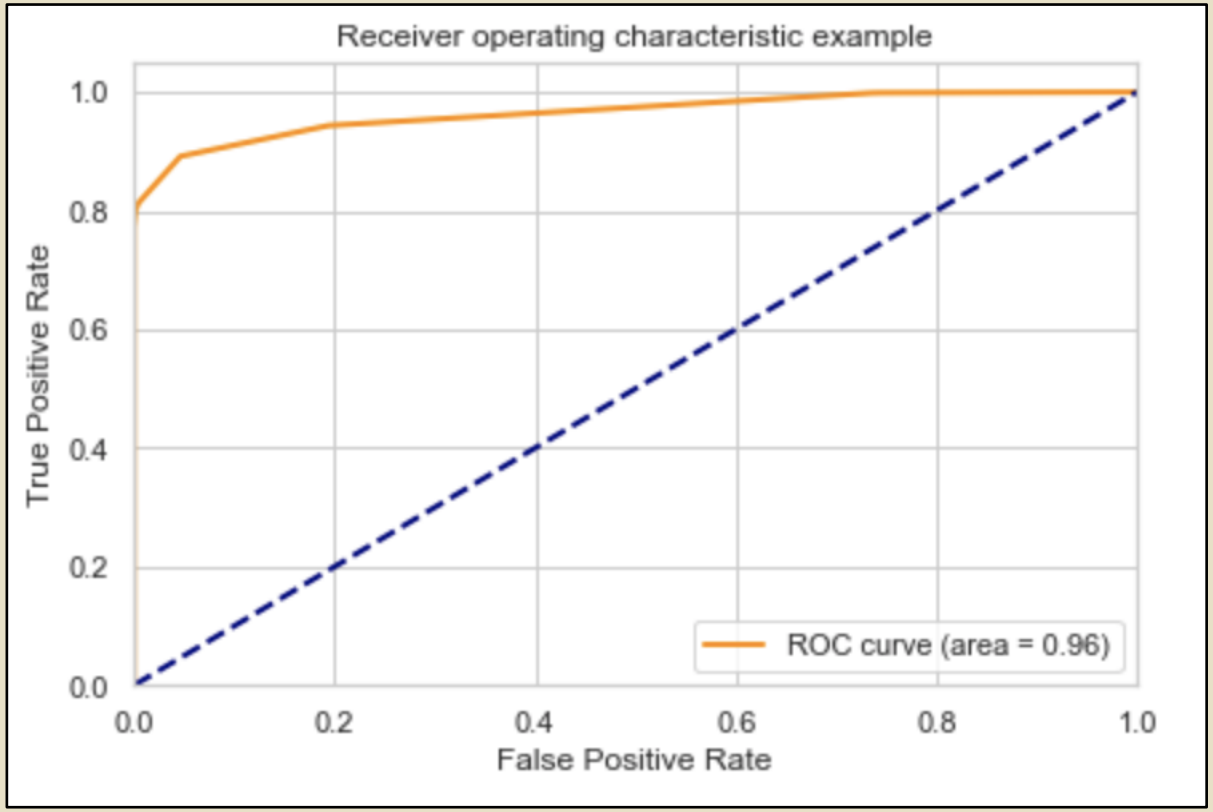


The plot above exhibits that if our "d" value equals any number from 8-50, our data will most likely become overfitted. Therefore, any value from 4-7 would work just fine for this model.

	accuracy	sensitivity	specificity	d
DT-1	0.882247	0.982478	0.781791	3
DT-2	0.945136	0.993538	0.896625	4
DT-3	0.952538	0.991674	0.913314	5
DT-4	0.957452	0.998260	0.916552	6
DT-5	0.957452	0.998260	0.916552	7
DT-6	0.968213	0.995402	0.940964	8
DT-7	0.968213	0.995402	0.940964	9
DT-8	0.968213	0.995402	0.940964	10
DT-9	0.968213	0.995402	0.940964	12
DT-10	0.968213	0.995402	0.940964	14
DT-11	0.968213	0.995402	0.940964	15
DT-12	0.968213	0.995402	0.940964	30
DT-13	0.968213	0.995402	0.940964	50

Accuracy of 94-95% predicts our model best

Decision Tree Model (2)



- After testing our model with both 74 and 10 variables separately, like in Logistic Regression, we realized that for this model, using only 10 works best.

- Using d=7, which is our limit as explained in our previous slide, our ROC curve displays that our model has an area of 0.96. Meaning that our model is 96% better than randomly predicted results.

Random Forest Model (1)

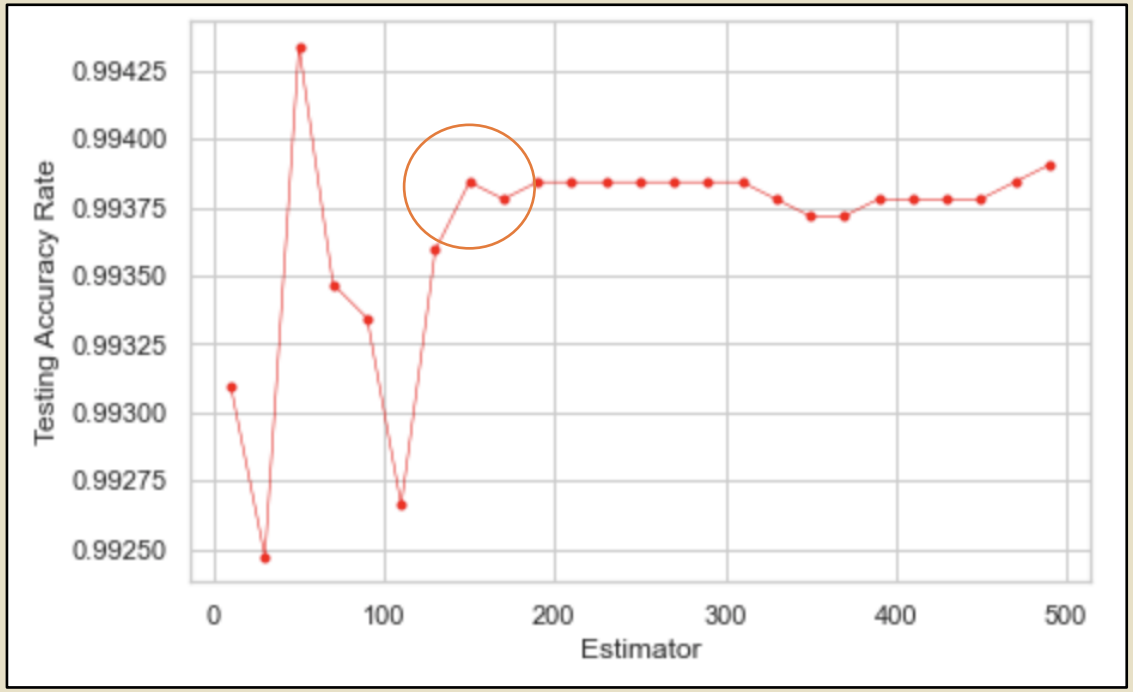
After testing both variable combinations with this model as well, we concluded that neither of these combinations could accurately predict our model. However, for our presentation we are showing the model with all 74 variables.

	accuracy	sensitivity	specificity	e
RF-8	0.993842	1.0	0.987670	150

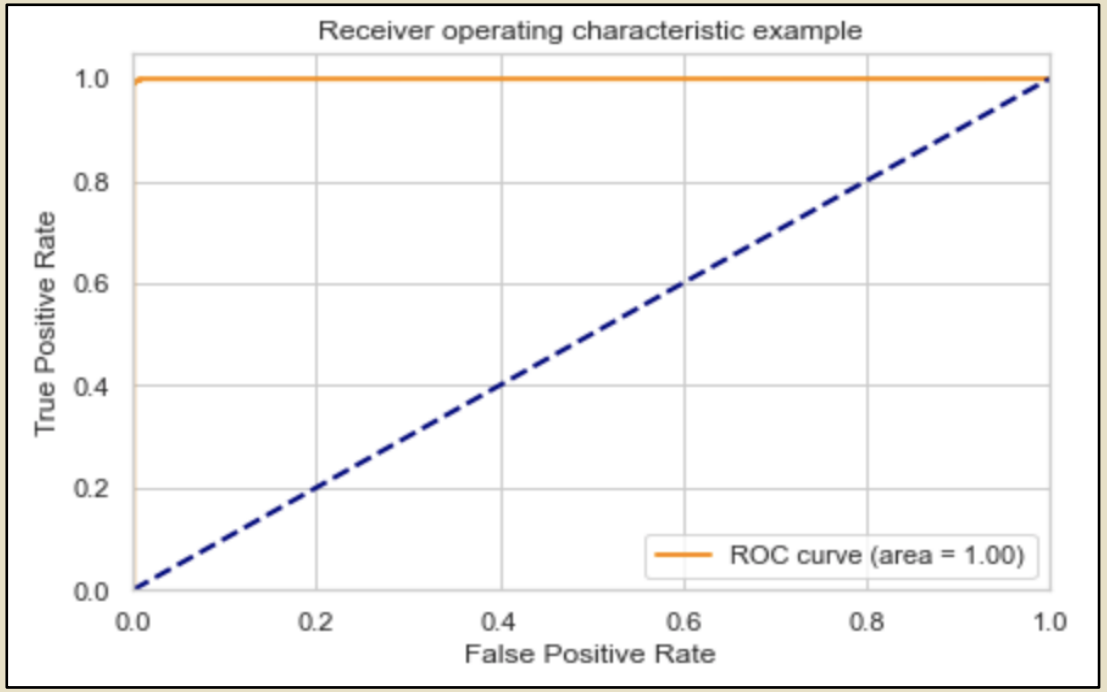
We can conclude that our model is overfitted due to:

- All values on sensitivity are equal to 1.
- Most estimators used have an accuracy of 99%. Hence, very uncommon.
- Values on specificity are too similar.
- There is not much variance, which may be due to repetition on the variables selected in our test and training sets.

Random Forest Model (2)



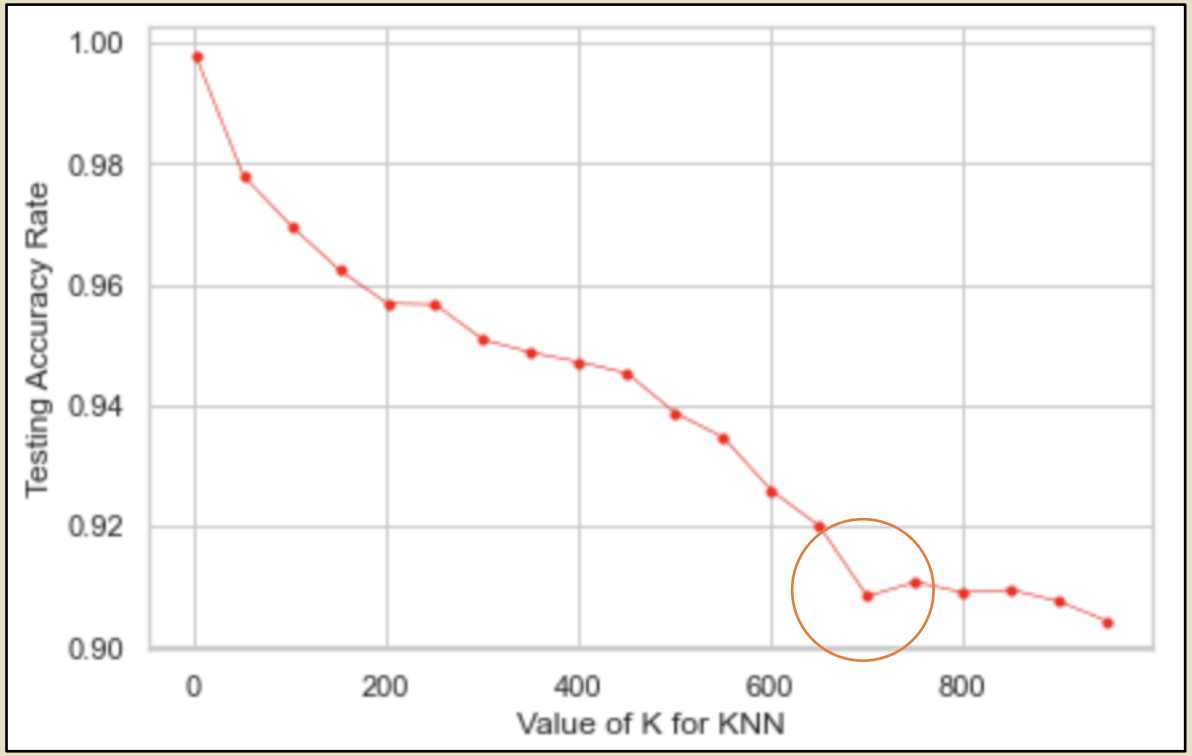
The plot above, exhibits the slight differences between the estimators chosen to test our model. We choose 150 as our number of estimator. However, none of which, give us accurate results.



Our ROC curve gives an accuracy of 100%. Technically, this tell us that we have perfectly predicted our model. However, having an accuracy this perfect is bizarre. Therefore, we concluded that our data is indeed overfitted and dropped this model from consideration.

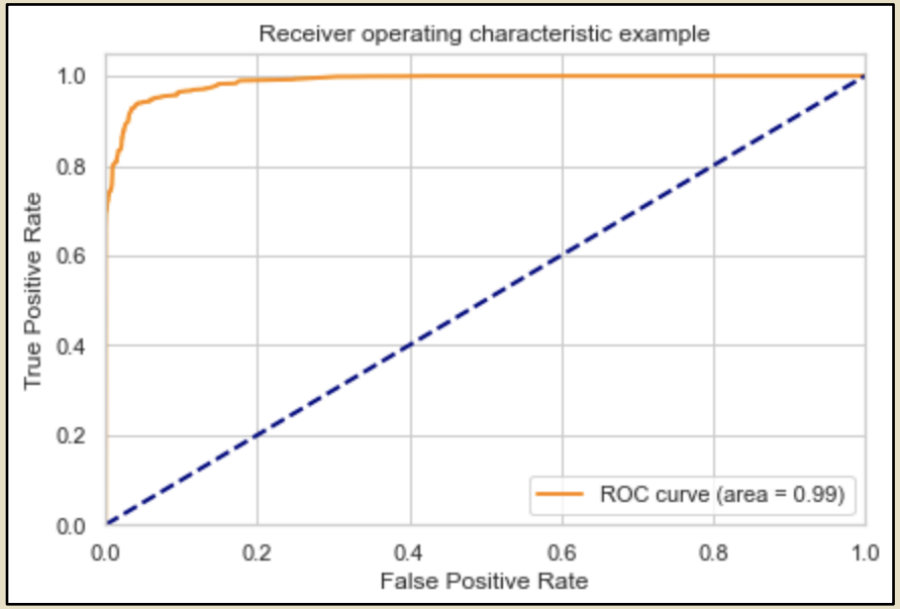
K-NN Model

Because there are 31612 observations, we chose the value of k in the range of 2 to 1000 as step of 50 to avoid overfitting which might be brought by small value of k.



	accuracy	sensitivity	specificity	k
KNN-15	0.908373	0.979744	0.836841	702

As we can see in the plot, the curve goes flat after k as 702. Therefore, we chose 702 as the final value of K and get the ROC curve as below.



PERFORMANCE & CONCLUSION

Performance Indicators | Final Results

- Here is the summary of performance indicators of all our models. For the former two models, we get the result based on 10 selected variables, while the last two were based on all variables. Our final goal is to find the model with maximized accuracy, sensitivity and specificity.
- For RF-8, the values are all too high to be real, which might because of overfitting.
- For other three models, we can see that DT-5 has the highest accuracy and balances the value of sensitivity and specificity very well.
- Therefore, DT-5 would be the best model for our dataset, which is ***Decision Trees with depth as 7.***

	accuracy	sensitivity	specificity	d	e	k	sensi.*speci.
Logistic Regression 2	0.918947	0.945321	0.892515	NaN	NaN	NaN	0.843713
DT-5	0.922742	0.953647	0.891767	7.0	NaN	NaN	0.850432
RF-8	0.993717	1.000000	0.987421	NaN	150.0	NaN	0.987421
KNN-15	0.908373	0.979744	0.836841	NaN	NaN	702.0	0.819890

Business Problems Conclusion

- ***What features should a product have in order to get good ratings?***
  - We believe that a product that has a combination of quality, support, and good fit would be a product that will most likely receive good ratings since after our analysis we realized that most women look for these three features for the most part.
  - Another feature women seem to prefer when purchasing underwear is comfort. Although it may be obvious, some brands focus a lot more on the design rather than the functionality of a product. We would suggest to have a balance on both design and functionality on products in order to receive better ratings and increase sales.

- ***Do women value more the quality of a product rather than its aesthetics when purchasing a product?***
  - Women seem to be more aware of the quality of underwear they purchase rather than its aesthetics. They look mostly for support and good fit when picking underwear, specifically bras. Hence, we can conclude that most women would pay a hire price for a bra that fits well and is made with good quality materials.
  - For the most part we have noticed that brands should focus a lot more on the "fit" rather the "design of bras", which is the number one problem for women that looking for a good bra. There is no doubt that if a woman finds a bra that fits well, adds support, is comfortable and well made, she will most likely become a loyal client.
  - Cotton fabrics are very popular for underwear pieces due to its breathability and non allergenic fibers. Natural fibers seem to be the materials of choice for underwear because they help control temperature and moisture, making them the best quality fabrics for underwear pieces. We have noticed that after our word cloud analysis, most brands seem to focus more on Nylon and Spandex fabrics, which are synthetic and not breathable. If we could suggest these brands an ideal change, it would be to op for natural fibers in order to increase client satisfaction and an increase in sales.

LIMITATIONS

Limitations

- We have very few observations coming from Victoria' Secret, which can cause our model to be a bit bias.
- Not all companies showcase information about both bras and panties. Hanky Panky's observations were purely made about panties and Victoria's secret only about bras. Therefore, decisions about either of those two, automatically drops one of those brands from consideration.
- Our dataset does not have a big amount of observations about "panties" as it does about "bras". Therefore, most of our result mostly narrows down information and conclusions about the impact of bra sales on Amazon.
- The dataset is just the summary of products. There is no data showing the selling with time change. Therefore, we can't evaluate the products by the actual behavior of customers and see the relationship between sells and ratings.
- Our ratings are only numerical, if our dataset had text, we could use a word cloud in order to narrow down the likes and dislikes about a specific product and be able to have a more specific conclusion about a good product vs a bad product in order to advice each brand on what to improve.
- Our dummy variables are made from the high frequency words in description and style attribute. Due to some useless words in word selection, some dummy variables don't make a lot sense, which makes our analysis not that rigorous.

THANK YOU