H&M Summer Retention Prediction

Group #3 - Yi-Hsuan Wu/Farid Freyha/Yu-Tung Chang/Nan Wang/Yu-Cheng Lai

Executive Summary

In this project, we focused on identifying whether customers would retain to H&M in September 2020 after the peak summer season from June to August 2020, and utilizing the information to figure out customers with what kind of characteristics would retain and products with what kind of features would be most likely purchased even after the peak summer season. After running multiple machine learning models and comparing their performances, we found that the neural network model performed the best, with 96% accuracy of identifying whether the customer will retain in the month after the summer season. Besides, according to the prediction result form neural network model, we found that for retaining customers, the biggest age group retained in September 2020 ranges from 16 to 25; as for product features, black colored items, items with solid patterns, items within women's collection and full body garments were preferred by retaining customers. In addition to the insights gained from the prediction model result, we recommend H&M data science team working more closely with the sales and design department to provide more data-backed insights(i.e. sales prediction and product recommendation system) for H&M when constructing customer retention strategies after the peak summer season.

Introduction

H&M Group is a family of eight fashion brands, established in 1974. The company has expanded globally throughout the years and reaches consumers across different segments. As customer acquisition and retention are becoming more relevant for the fashion industry and with the digitalization trend in the industry, the company is putting more and more attention to its online platform. In order to keep its competitive place in the market, H&M Group has to better strategize their planning and marketing that help increase its customer retention rate, by analyzing the data from customer previous transactions.

According to H&M customer transaction data from September 2018 to September 2020, we found that sales peaks appeared every summer season from June to August(*Graph 1*), which accounted for 30.29% of total transactions in one year. We would like to know whether customers will retain after the summer sales season, which means the customer will at least make one purchase in September 2020. Thus, we could drive the insights from the result of machine learning prediction to identify the important indicators of loyal customer groups, the popular items they were interested in, and uncover the purchasing patterns from those retaining customers.

Exploratory Data Analysis

In order to better understand the datasets we are using in the prediction model, we first conducted exploratory data analysis respectively on customer datasets, product datasets and transaction datasets. We found that the average age of customers is 30 years old. While the age group between 21 to 30

accounted for 43.5% of total transactions volume, the age group between 31 to 40 only accounted for 16.7%. Among all customers, 55% of them are active in the membership club whereas 45% of the customers have left the club or have registration records in the club. For the sales channel, 65% of sales came from online sales while 35% were in-store sales. From product datasets, we found that Ladieswear has 65% of total sales across all sections, including Menswear, Sports and Baby/Child sections. The color of Black is the most popular color, which accounts for nearly 50% of total transactions. The top three popular product categories are women's everyday collection, women swimwear, and women underwear. For transactions, we found that 25% of sales happened on weekends in the summer season from June to August in 2020.

Data Processing

Subset Creation

We have three datasets including customer, product and transaction. The customer datasets include the information of customer age, membership club status, fashion news frequency, etc. The datasets of the product provided name, category, type, color, graphical appearance, garment group of the product. The transaction data were given the purchase history of customers across September 2018 to September 2020, along with supporting information of price, sales channel, customer ID, and product ID.

Our first step is to extract the H&M Group transaction data from June to September in 2020, and merge the customer datasets and product datasets on customer ID and product ID respectively. Next, we aggregated the merging dataset to the customer level, creating a peak summer season subset by adding new variables organized from the merging dataset. Also, we removed customers that had purchases after August in order to identify whether customers who had purchases in the peak summer season (June to August) would retain to H&M in September. Then, we created a random subset that is 10% from the peak summer season subset so as to perform model predictions more efficiently.

Feature Engineering

For the new variables, we removed those price outliers and came up with new dummy variables based on whether customers made the purchase on weekends, did they subscribe to the H&M promotion news and where did they make their purchases, online or in stores. We also created a new group: whether they are active in the membership club or they have already left the membership club. In addition, we created categorical variables for the product features such as product categories, colors, textures, graphic appearances and cloth styles.

Feature Selection

In order to decide the variables to be put in the models, we calculated the correlations between the independent variables and the dependent variable, the retention of September. We decided to choose variables sharing the correlation larger than 0.14, so the last variable we remained would be the customer group with ages ranging from 30 to 46.

Before running the machine learning models, we split the dataset into training, and testing samples with 80% of datasets for training and 20% for testing. At the same time, we checked the proportion

between the retention and no retention in the random subset, training samples and testing samples in order to examine whether the imbalance problem in datasets is significant or not. After checking, we found that percentages for retention and no retention are 75% and 25% respectively, showing that the imbalance problem for datasets is not significant. Then, we created validation samples and made sure it didn't include the sample from the training dataset so as to avoid the risk of leakage while performing model prediction.

Prediction Models

Logistic regression & Stepwise Regression

Since our task is to predict the binary outcome of customer conversion retention, several classification algorithms were involved in our cases. We started with the simple supervised machine learning model: Logistic Regression. Considering there are multiple variables, we chose to use "stepwise" regression modeling since it improves the model generalizability by reducing the number of variables. We started with all variables first and then removed one variable at a time until we got the " best" model.

Based on the result of stepwise regression, we should choose the model with the lowest AIC value. The "best" model includes variables from the training dataset, such as the dummy variables of different colors, the master of the color, whether the product is ladies wear, number of purchases for June, July and August, total amount spent on purchase, the product price in August, and all four age groups but not limited to the ones mentioned above.

Our best predicted logistic model based on stepwise regression has a sensitivity score of 0.7814. Which means that about 78% of true positives (e.g. the customer will at least purchase once in September 2020) the model can predict. To better evaluate the model, we also looked at other model assessment criteria, like, accuracy and AUC score (i.e. the measure of the ability of a classifier to distinguish between classes). The model has an accuracy score of 0.9359, which means it correctly identified 94% of the two classes. To be noticed, the result we got from the original logistic regression model and the one after stepwise did not show that much difference, indicating the correlations between independent variables and the dependent variable are useful for machine learning for this dataset.

K-Nearest Neighbors(KNN) Algorithm

As the KNN algorithm is commonly used for its ease of interpretation and application of classification problems, we also tried the KNN model. The KNN algorithm helps to find the closest component with shared attributes; that is, we predict the category of the test point from the available class labels by finding the Euclidean distance between the test point and trained k nearest feature values, which is useful for target segmentation. In this case, we used hyperparameter tuning to determine the optimal value of K, searching the value of K ranging from 1 to 30. Eventually, we striked the balance between training accuracy and testing accuracy and found that 3 is the best value for K and received the sensitivity of only 0.197 with accuracy of 0.75 and AUC score (i.e. the measure of the ability of a classifier to distinguish between classes) of 0.52, which means the classifier is not able to distinguish between Positive and Negative class points.

Support Vector Machine(SVM) & Naive Bayes

The SVM and the Naive Bayes models are also popular classification machine learning models that have been commonly used. The SVM model separates two classes of data points to find a maximum distance between data points of both classes. Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, which is a mathematical formula used for calculating conditional probabilities. However, these two models did not work as well as other models, each ending up with the result of accuracy of 0.788 and 0.77, AUC scores of 0.58 and 0.5927, and sensitivity of 0.188 and 0.57.

C5.0 Decision tree & Classification tree

We ran the model with C5.0 decision tree and it had a sensitivity of 0.8407 when compared to the sensitivity of the stepwise regression model at 0.7814. We also ran the classification tree, ending up with a sensitivity of 0.7202, lower than the result we got from both the C5.0 decision tree and the stepwise model. We further pruned the tree by specifying the Mallow's Cp to 0.00032068 since at this Cp level, the value of cross error were minimized. For the result, we got a sensitivity of 0.7008, which is lower than the original tree. However, the accuracy is 0.9072, slightly higher than 0.9017 of the original tree, showing the pruned tree performed better on predicting the true negatives.

Random Forest

For the Random Forest model, we ran it on Python. Before we start, in order to find the best hyperparameters, we use the hyperparameter grid to search for the best hyperparameters for the model. According to the result of the hyperparameter grid, we set bootstrap sampling method, the number of trees to 100, the max depth of each tree to 82, the minimum number of samples required to split an internal node to 2, the minimum number of samples required to be at a leaf node to 2. From the prediction performance, we could see an accuracy score of 0.893, a sensitivity of 0.622 and an AUC score of 0.8. When comparing with decision tree, stepwise regression and random forest, we noted that both C5.0 and rpart decision tree models had higher accuracy and higher sensitivity than the random forest model.

Gradient Boosting Machine(GBM) and XGBoost

Gradient boosting refers to a methodology in machine learning algorithm where an ensemble of weak learners is used to improve the model performance in terms of efficiency, accuracy, and interpretability. Similar to Random Forests, Gradient Boosting is an ensemble learner, and XGBoost is the optimized gradient boosting algorithm. The key difference between GBM and XGBoost is that XGBoost focuses more on regularized model formalization to better avoid overfitting, feasible to deal with large datasets, and usually ends up with better results.

The result from two algorithms shows that XGBoost has much better prediction performance than gradient boosted decision trees did. While GBM has an accuracy rate of 0.8 and AUC score of 0.64, its sensitivity is only 0.32. Nevertheless, for XGBoost, it is the best model among all decision tree models with 97% accuracy, 94% ability to classify customer retention and no retention and sensitivity of 0.89.

Artificial Neural Networks

Artificial Neural Networks (ANN) is based on brain function and is used to model complicated patterns, and it is a deep learning model that is based on the concept of human brain neural network. It is composed of a large number of highly interconnected processing elements known as neurons to solve problems. When we ran the Neural Network model in R, we got an accuracy score of 0.9767, a sensitivity score of 0.92, and an AUC score of 0.91. When compared to other models, we can notice a significantly higher difference across all four metrics mentioned above.

Model Evaluation

After running all the models on validation samples(Table 1), we found that the neural network model has the best prediction performance of highest accuracy, sensitivity, and AUC score for this dataset. Since the neural network has the best result for the validation sample, we then ran the Neural Network model again on the testing sample. Under its prediction performance(Table 2), we got 97% accuracy to distinguish whether the customer will retain or not, and 87% sensitivity to identify the customer who would actually make at least one purchase at H&M in September 2020.

Recommendation

Technical Recommendation

As mentioned above, there are three indicators we look at while evaluating a model. We believe sensitivity is the most relevant for H&M as it measures how well a machine learning model can detect true positives. It is also known as the true positive rate (TPR) or recall. It evaluates model performance by emphasizing on the number of positive instances the model was able to correctly identify. In our project, we are interested in the retention of customers, more specifically, the ones who will remain as customers and make purchases in September. Thus, sensitivity is the indicator we should focus on since it can assist us with evaluating how well the model can identify how likely customers who will retain actually retain in the next period.

With an accurate model, the data science team could further focus on the customers the model predicted would retain and recommend new products or services they may be interested in by tracking their purchasing habits before so they would not only make a purchase in the following months but couple or at least purchase more products. As for the customers that weren't retained, we should also try to figure out the important causes of their churn rate, whether it being the price point, the quality or the lack of specific availability that these customers were looking for.

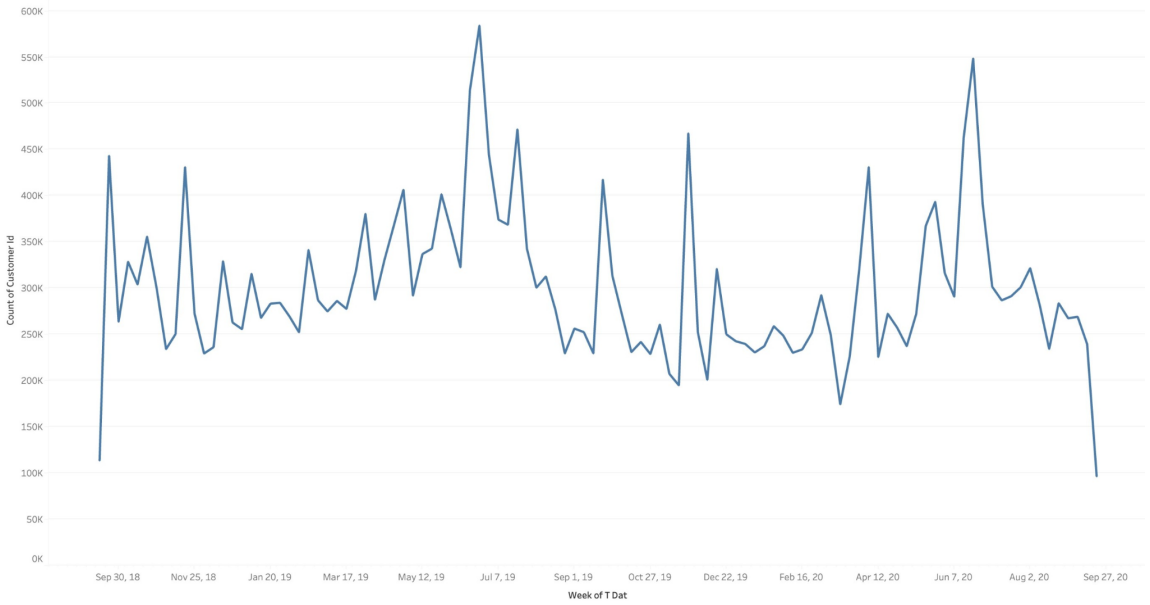
Managerial Recommendation

The Neural Network model mimics the consumer buying decision process, which helps us discover insights in the customer journey and has led to multiple insights. We found that the biggest age group retained in September 2020 ranges from 16 to 25 at 28.04%, followed by the group range from 25 to 30 at 25.26%(Graph 4). We also recommend H&M to focus on the color as 45.76% of our retained customers purchased black colored items, followed by white at 21.12%(Graph 5). In addition, for the graphical, retained customers prefer solid at a majority rate of 85.94% when compared to all over

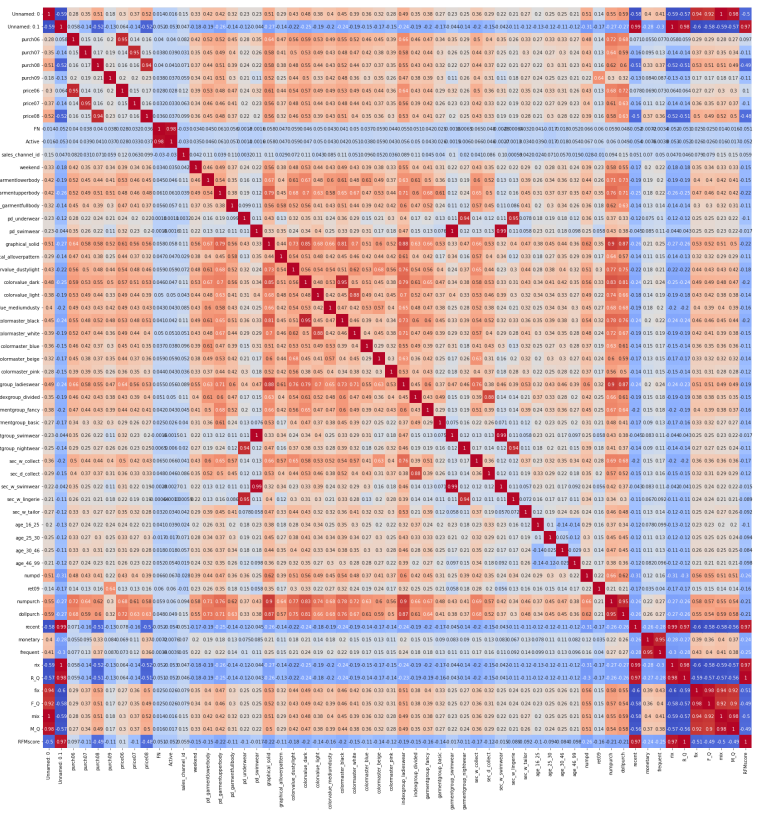
pattern at 14.06%(Graph 6). Another distinction we found was within sections where women's collection accounted for 38.18%, followed by the divided section's collection(general neutral) of 20.57%(Graph 7). We also noticed a high demand for upper body garments, such as t-shirts and shirts, compared to other garments accounting for 52.04%(Graph 8), However, considering that September marks the end of summer and the beginning of the fall season. We expect a change in the percentage of sales by product category and color due to season changes and trends, so we strongly suggest working more closely with the sales and design department, as providing them with more data-backed insights will assist the teams to better predict sales and plans for the upcoming season, especially as retained customers are more likely to maintain their past purchase patterns and purchase more items from those specific categories. As well as, advance H&M's product recommendation system to enhance the personalized shopping experience.

Appendix:

Graph 1: H&M Group Transactions across September 2018 to September 2020



Graph 2: Correlations between independent variables and dependent variable



Graph 3: KNN: Determining the Optimal Number of Neighbors

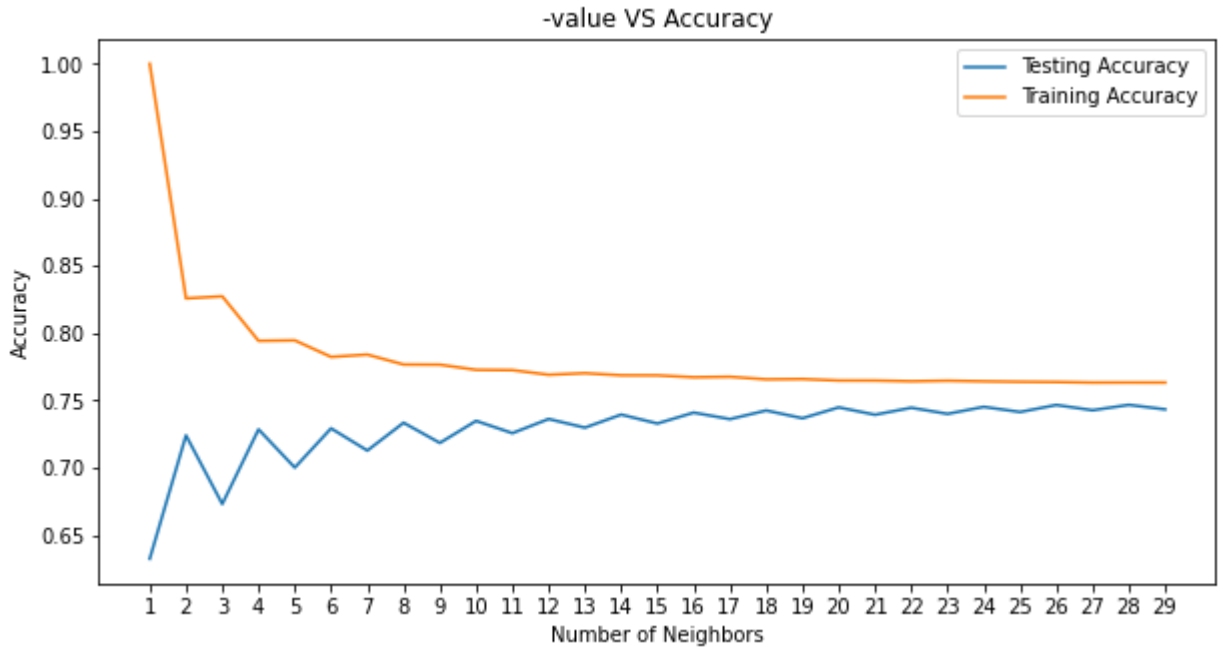


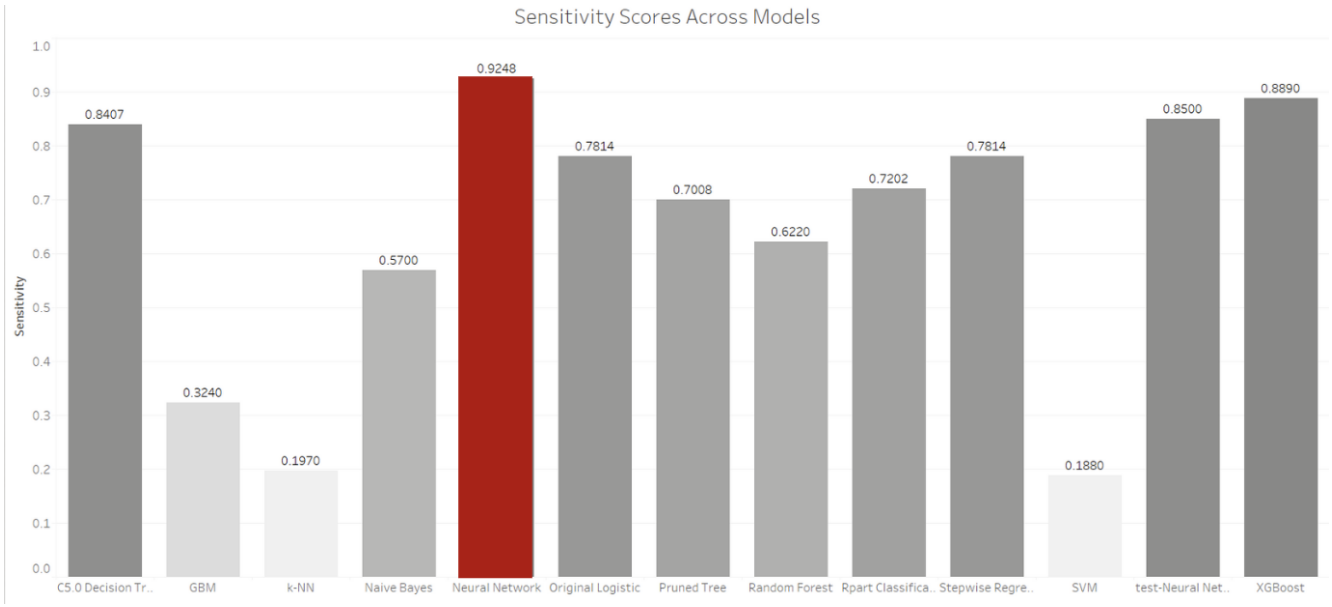
Table 1: Model Prediction Performance across Different Models (Validation Sample)

Model/Performance	**Accuracy**	**Sensitivity**	**AUC score**
Original Logistic	0.936	0.7814	0.9591
Stepwise Regression	0.9359	0.7814	0.9591
k-NN	0.69	0.197	0.5235
Naive Bayes	0.77	0.57	0.5927
SVM	0.788	0.188	0.58
Rpart Classification Tree	0.9017	0.7202	0.8396
Pruned Tree	0.9072	0.7008	0.8366
C5.0 Decision Tree	0.945	0.8407	0.9093
Random Forest	0.893	0.622	0.80
GBM	0.8061	0.32403	0.6411
XGBoost	0.967	0.889	0.94
Neural Network	0.9767	0.9248	0.9589

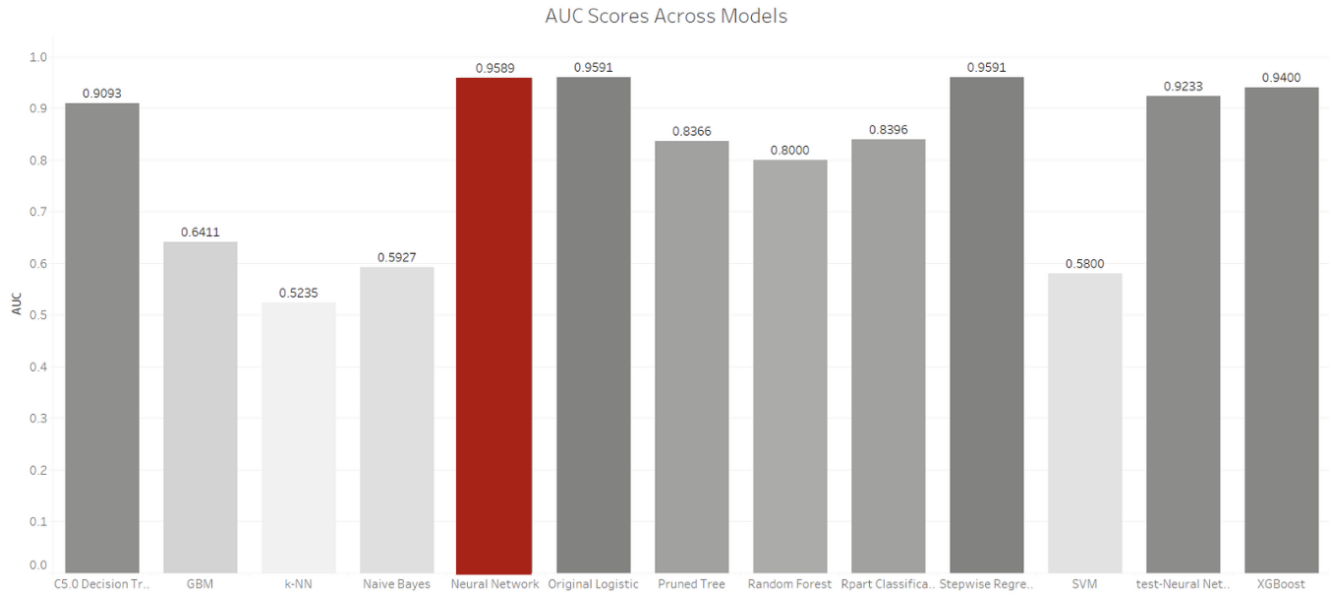
Table 2: Prediction Performance in Neural Networks Algorithm and Logistic Regression with Testing Samples

Model/Performance	Accuracy	Sensitivity	ROC
Logistic	0.9354	0.7783	0.9493
Neural Network	0.9616	0.8500	0.9233

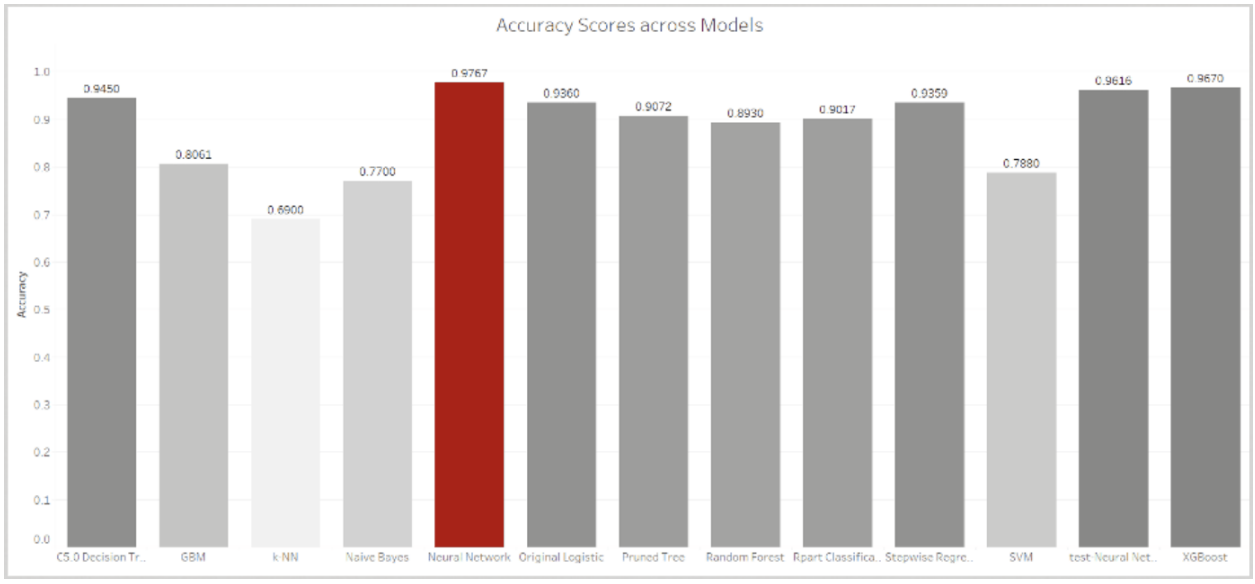
Graph 4: Sensitivity Scores Across Models



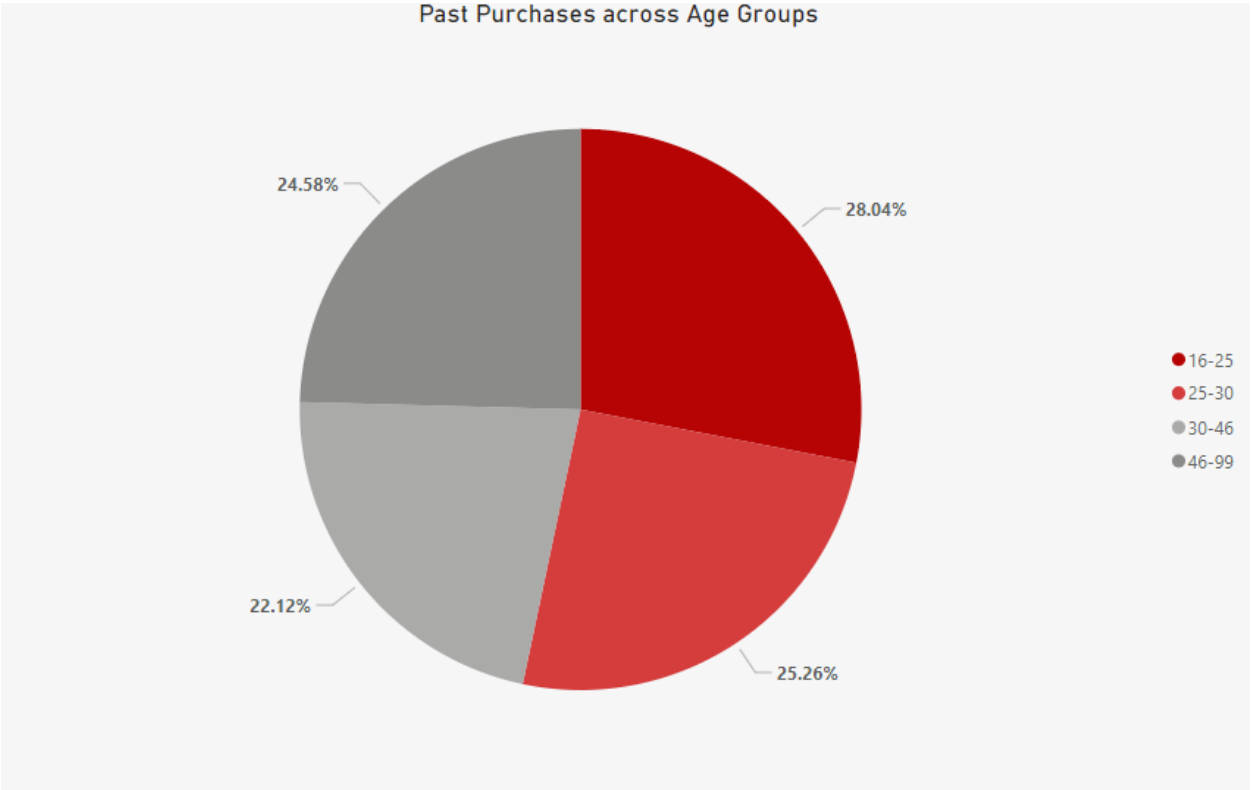
Graph 5: AUC Scores Across Models



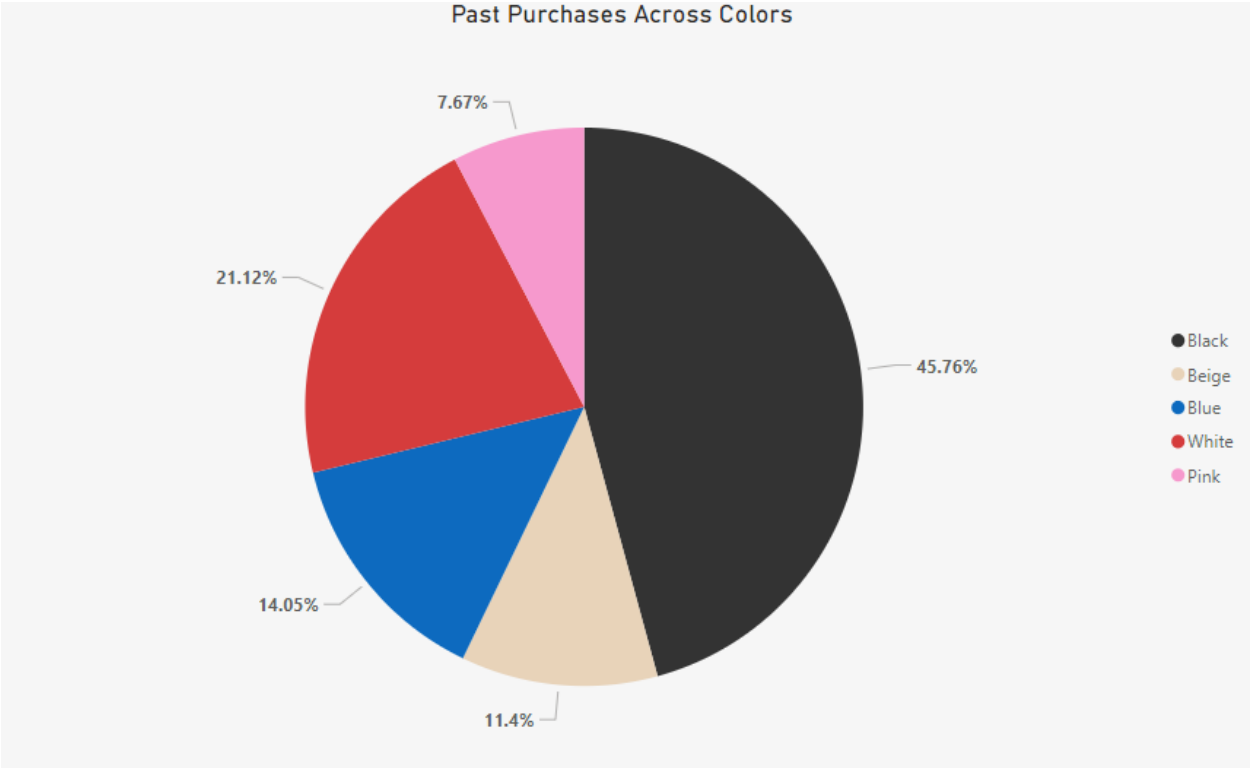
Graph 6: Accuracy Scores Across Models



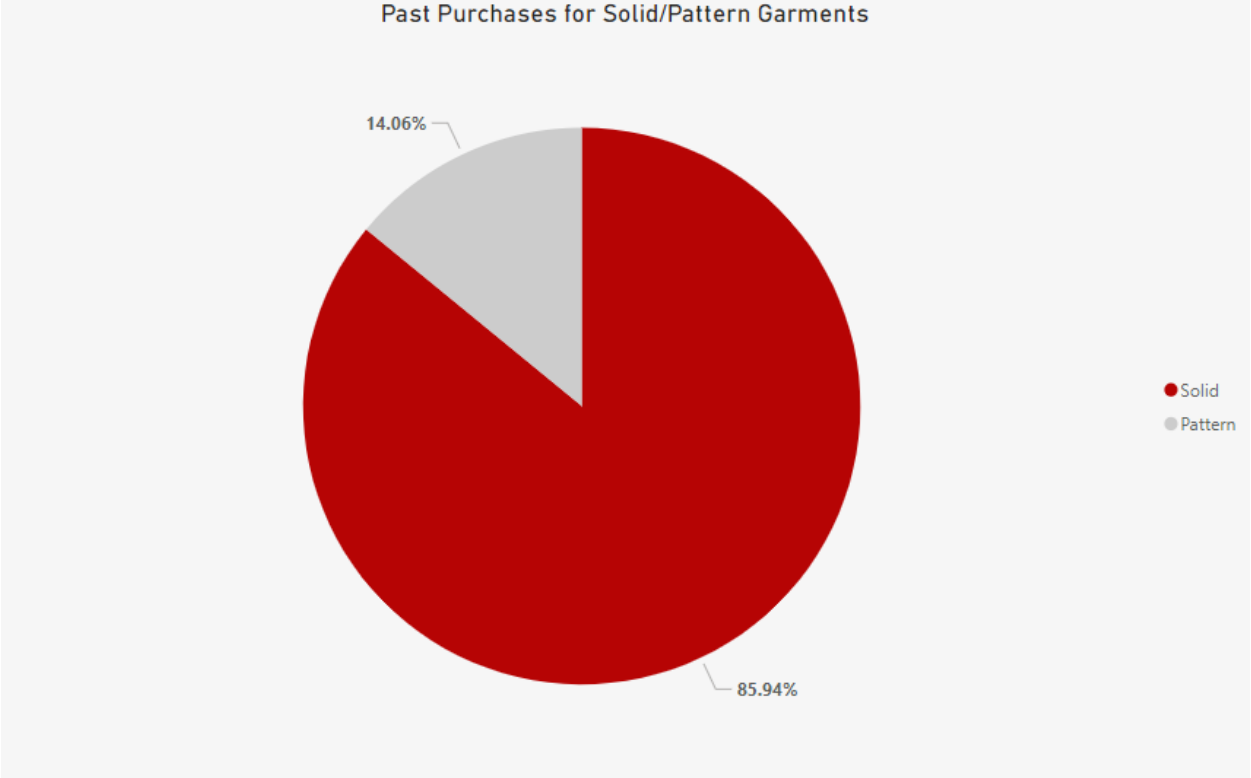
Graph 7: Past Purchases across Age Groups



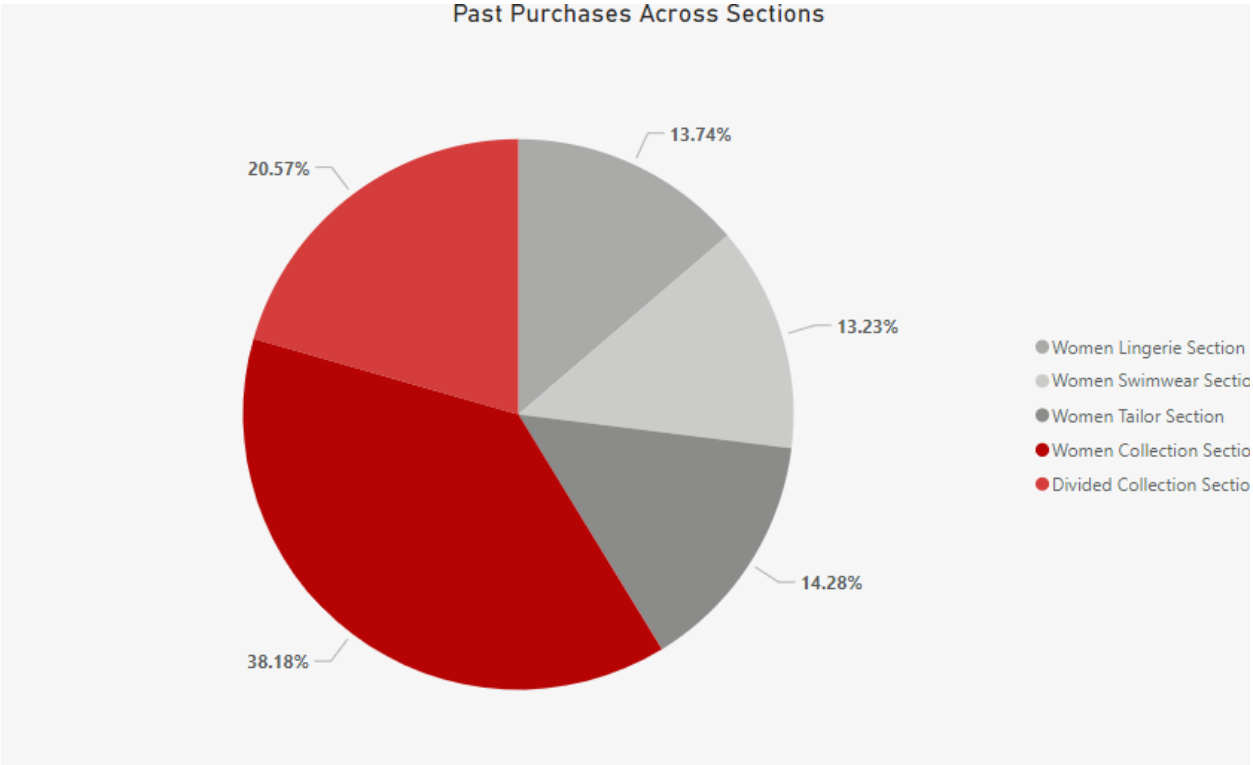
Graph 8: Past Purchases Across Colors



Graph 9: Past Purchases for Solid/Pattern Garments



Graph 10: Past Purchases Across Sections



Graph 11: Past Purchases Across Garment Types