Module 15 | Final Project

Morris Wu

2024-12-03

Data Set 1 – Pricing cars

The supervised learning method I selected for the cars_big dataset is "random forest," which combines the predictions of multiple decision trees. To fit the model with regression, I followed five main steps: First, I processed the data by replacing unspecified values and converting categorical values into numerical or factor types. Second, I ran a shallow random forest to identify columns with low contribution to the predictive value and removed them from the dataset. Third, I split the data into training and testing sets. Fourth, I imputed missing values in the dataset using the na.roughfix function, which replaces missing values with the median or mode. Finally, I trained the model using the training dataset, setting price as the target variable.

Data Processing

Details on how I processed the data set:
• Replaced missing values and "unsp" in columns with NA.
• Converted Boolean value columns, such as isOneOwner, to numeric values where 1 equals TRUE and 0 equals FALSE.
• Removed irrelevant characters in certain columns, such as removing "L" from displacement and whitespace from wheelSize.
• Converted all columns with numerical data to numeric types and all columns with categorical data to factor types.

```
##                   %IncMSE IncNodePurity
## ...1         -1.59985771    12873371.8
## trim          0.00000000           0.0
## subTrim       0.00000000           0.0
## condition     5.36972292    27311296.9
## isOneOwner   -0.63987499     1130347.5
## mileage      -3.08201164    36285967.1
## year         -2.42001761     6881725.1
## color         6.95198663    57585949.1
## displacement  0.00000000           0.0
## fuel          0.00000000           0.0
## state         0.65247712    47177383.0
## region       -1.00557001    28442051.6
## soundSystem  -1.17576387     1704850.1
## wheelType     0.00000000      807516.2
## wheelSize     0.04377423     6047776.9
## featureCount  2.66704388    26607964.0
```

Details on feature Selection Process:
• First, I copied the processed dataset and removed rows with missing values using the na.omit function to ensure the dataset was clean and ready for modeling.
• Second, I trained the copied dataset using the Random Forest model, limiting the number of decision trees to 500. This was done to reduce processing time while maintaining accuracy.
• Third, I used the importance function to evaluate the importance of each feature.

Based on the output of the importance function, I focused on features with lower to negative %IncMSE values, as these features either contributed little to the model or negatively impacted its performance. However, considering that the sample size was significantly reduced due to the na.omit function, I opted to remove only 4 features to avoid over-simplifying the model which includes: "...1", "subTrim", "isOneOwner", "region".

Fitting the Model

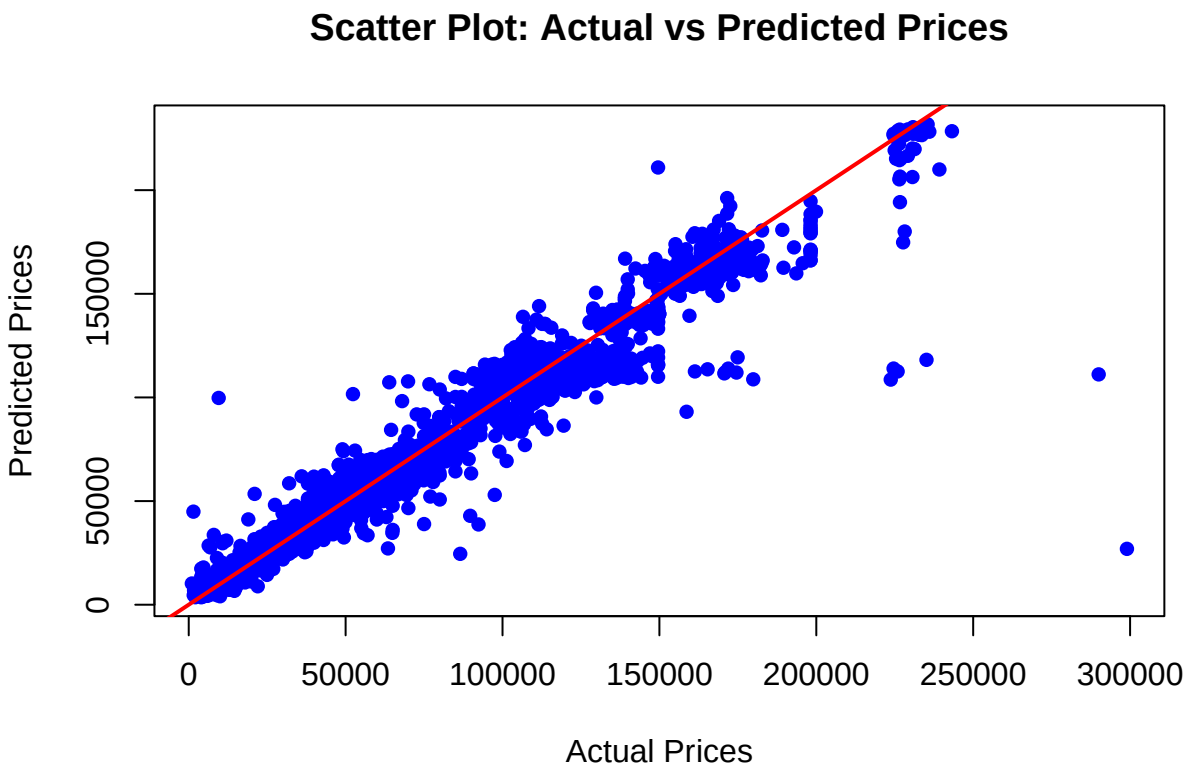
To fit the models and run the regression, I first generate a random sample and split the dataset into 70% training and 30% testing datasets. I then handle the missing values using na.roughfix to ensure there are no missing values. After that, I build the Random Forest regression model to predict the price variable. Due to the large dataset size, I specify the model to use 300 decision trees and limit the number of features randomly selected at each split to 5. Finally, I generate predictions for the price variable in the test dataset using the trained model. By comparing the predicted values with the actual results, we can evaluate the accuracy of the model.

Result summary

```
## MAE: 4539.246
```

```
## RMSE: 8013.091
```

```
## R-squared: 0.9684273
```



Based on the outcome, my model has a Mean Absolute Error of 4412.012, meaning the model's predictions are off by approximately $4,412 from the actual prices. Additionally, the model has a Root Mean Squared Error of 7089.5, which indicates that some predictions deviate significantly from the actual values. The R-squared value of 0.9748817 is a good sign, showing that 97.49% of the variability in price is captured by the Random Forest model.

By looking at the scatter plot of actual prices versus predicted prices, we can also conclude that the model captures the trend of car prices since the trend line aligns closely with the scatter plot points. Although the model is fairly accurate, there are still some outliers and room for improvement. For example, in the plot, some cars are actually priced at 20,000+ dollars while the model predicts them to be 10,000 dollars cars. This indicates that the model does not perform as well, especially when the actual prices of cars are on the higher side.

Data Set 2 – Market segmentation

Supervised model are mainly used for predicting dependent variables with the independent variable. Meanwhile, unsupervised learn are used to find patterns and correlation between different factors. Since the project's objective is focusing on "identifying" market segments, it is more suitable to use unsupervised learning in this situation. If the question is asking us to predict how many tweet a followers have that are related to a certain broad area of interest, then supervised model will be recommended.

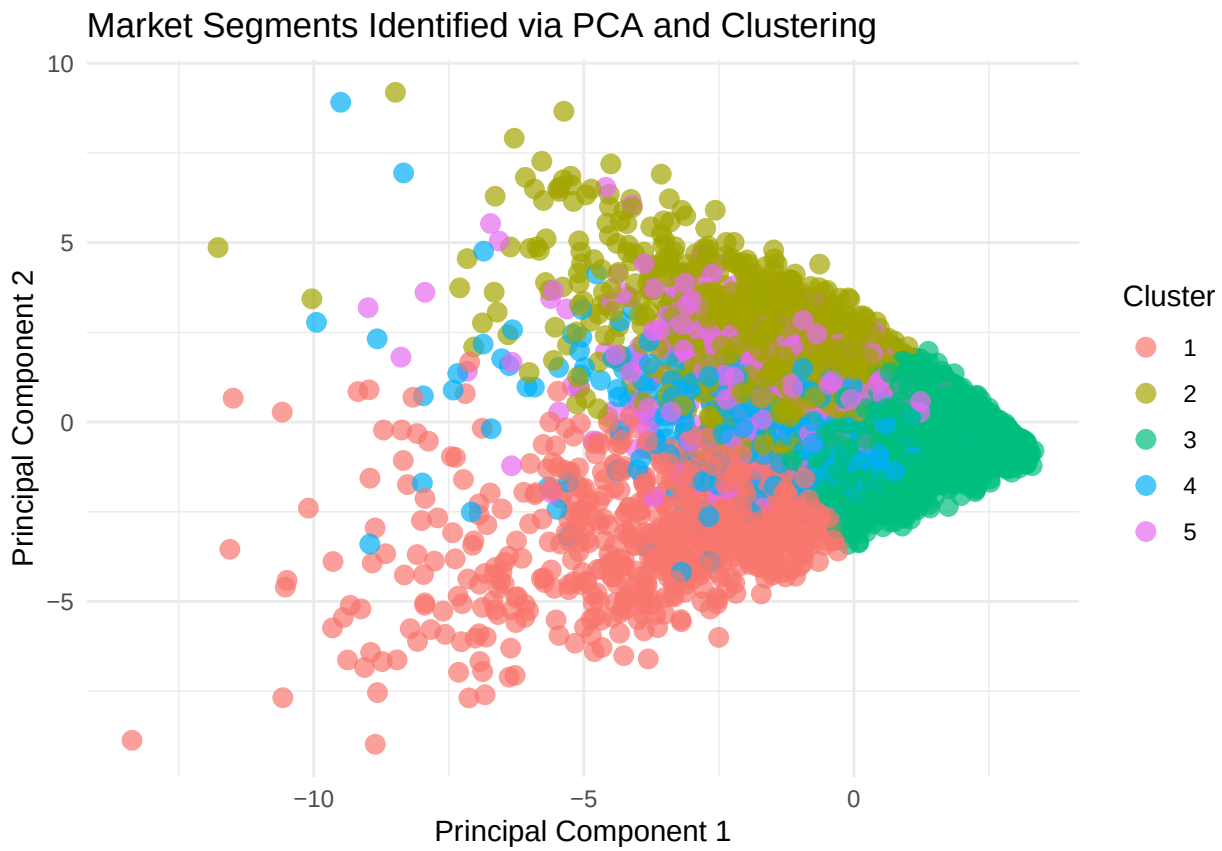
To process the data, I first removed the first column, and convert all variables into numerical values. Since the goal for the project is to identify the "market" segments for NutrientH20 based on their social media audiences, the customers' id does not provide contribution to our model. The next step I scaled to data ensuring that all variable have the same performances, and applied the PCA to the scaled data. Based on the Cumulative Proportion from summary pca results, we can see that we need the first 25 principal data

to capture most variance(90.6%) in the data. After selecting the first 25 PCs, I used the K clustering means to perform on the 25 PCs and created a total of 5 clusters.

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
## Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405
## Proportion of Variance 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476
## Cumulative Proportion  0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.93297 0.91698 0.9020 0.85869 0.83466 0.80544 0.75311
## Proportion of Variance 0.02418 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575
## Cumulative Proportion  0.73643 0.75979 0.7824 0.80287 0.82222 0.84024 0.85599
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513
## Proportion of Variance 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051
## Cumulative Proportion  0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911
##                           PC29    PC30    PC31   PC32    PC33    PC34    PC35
## Standard deviation     0.60167 0.59424 0.58683 0.5498 0.48442 0.47576 0.43757
## Proportion of Variance 0.01006 0.00981 0.00957 0.0084 0.00652 0.00629 0.00532
## Cumulative Proportion  0.94917 0.95898 0.96854 0.9769 0.98346 0.98974 0.99506
##                           PC36
## Standard deviation     0.42165
## Proportion of Variance 0.00494
## Cumulative Proportion  1.00000
```

Results Summary

```
## # A tibble: 5 x 26
##   Cluster    PC1    PC2     PC3    PC4     PC5      PC6     PC7     PC8     PC9
##   <fct>    <dbl>  <dbl>   <dbl>  <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
## 1 1       -3.48  -3.03   0.954   0.741 -0.230   0.0463   0.0160  0.303   0.108
## 2 2       -1.32   1.93   0.0923  1.30  -0.101   0.0873   0.200  -0.129  -0.0658
## 3 3        1.42  -0.303  0.0680  0.219 -0.0638 -0.00814  0.0332  0.0123 -0.0402
## 4 4       -1.19  -0.168 -3.66   -1.31  -0.352  -0.789   -0.412  -0.262   0.0939
## 5 5       -0.758  0.988  1.50   -2.81   0.934   0.456   -0.188   0.0961  0.138
## # i 16 more variables: PC10 <dbl>, PC11 <dbl>, PC12 <dbl>, PC13 <dbl>,
## #   PC14 <dbl>, PC15 <dbl>, PC16 <dbl>, PC17 <dbl>, PC18 <dbl>, PC19 <dbl>,
## #   PC20 <dbl>, PC21 <dbl>, PC22 <dbl>, PC23 <dbl>, PC24 <dbl>, PC25 <dbl>
```




```
## $PC1
##                     Variable    Loading
## religion           religion -0.2971000
## food                   food -0.2969095
## parenting         parenting -0.2940041
## sports_fandom sports_fandom -0.2877318
## school               school -0.2806379
##
## $PC2
##                     Variable    Loading
## sports_fandom sports_fandom -0.3169236
## religion           religion -0.3161528
## cooking             cooking  0.3142880
## photo_sharing photo_sharing  0.3030776
## parenting         parenting -0.2950822
##
## $PC3
##                          Variable    Loading
## politics                 politics -0.4899027
## travel                     travel -0.4242597
## computers               computers -0.3670315
## news                         news -0.3360356
## health_nutrition health_nutrition  0.2255148
##
## $PC4
##                          Variable    Loading
```

```
## health_nutrition health_nutrition -0.4634666
## personal_fitness personal_fitness -0.4444448
## outdoors                 outdoors -0.4147432
## college_uni           college_uni  0.2555873
## online_gaming       online_gaming  0.2207630
##
## $PC5
##                      Variable    Loading
## college_uni       college_uni  0.4870183
## online_gaming   online_gaming  0.4764231
## sports_playing sports_playing  0.3706463
## photo_sharing   photo_sharing -0.2296606
## tv_film               tv_film  0.2102380
```

After the 5 clusters were differentiated, I created a scatter plot to visualize the distribution of data points across the first two principal components, as they capture the majority of the variance in the data. The plot allows us to easily identify each market segment. Based on the visualization, clusters 4 and 5 appear more distinct, while clusters 1, 2 and 3 show significant overlap, which may indicate that they share more characteristics. Additionally, clusters 2, 3 and 4 are more spread out compared to cluster 5, which suggests that their followers may have broader and more diverse interests. In contrast, cluster 5 is more compact, indicating that its followers likely have a more niche interest.

To better understand what each cluster represents, the next step is to group the data by cluster, analyze the PCA scores for each cluster, and identify the features most strongly associated with each principal component. Additionally, I created a summary of the top features with the highest loading for the first five principal components to provide further insight.

As shown in the cluster summary(shows each cluster's PCA scores):
- Cluster 1 has strongly negative values in PC3 and PC5, with a positive value in PC2.
- Cluster 2 has strong positive values in PC2 and PC3.
- Cluster 3 has a strongly negative value in PC4 and a strong positive value in PC2.
- Cluster 4 has strongly negative values in PC1 and PC2.
- Finally, Cluster 5 has a slightly positive value in PC1.

As shown in the loading summaries (which display the loading of each principal component):
- The variance captured by PC1 appears to be more related to lifestyle, as features such as food, sports, and religion have higher loading.
- The variance captured by PC2 seems to be associated with health, as features like health_nutrition, personal_fitness, and cooking have stronger loading.
- The variance captured by PC3 indicates a contrast, where health is negatively associated with education and gaming.
- The variance captured by PC4 might relate to popular news, with features such as politics, news, fashion, and automotive having higher loading.
- The variance captured by PC5 reflects a trade-off between education and gaming on one side, and social interests (news, politics, and fashion) on the other, as suggested by the loading.

By combining the observations made from the two summaries, we can conclude that:
- Cluster 1 might represent groups of people who focus more on health but are less engaged in socially related topics such as news, politics, and fashion.
- Cluster 2 might represent groups of people who also focus on health but are highly engaged in education and gaming, showing a broader interest balance.
- Cluster 3 is very similar to Cluster 1, representing groups of people who prioritize health while being less socially engaged.
- Cluster 4 might represent groups of people who are less interested in both lifestyle and health-related topics, indicating limited engagement in these areas.

• Cluster 5 represents groups of people who have less interest in health but display more diverse interests in topics such as food, sports, and religion.

Since Clusters 1, 2, and 3 share high similarity and show overlap on the scatter plot, they can be combined into a single cluster. This consolidation simplifies the segmentation.

Conclusion:

In conclusion, NutrientH20's social-media audience can be defined into three market segments: people who focus on health, people who focus on lifestyle, and people who care less about both topics. People within the health-focused segment are likely to create posts related to topics such as nutrition, fitness, and cooking. Meanwhile, individuals in the lifestyle segment tend to post about food, sports, or their religious interests. The last group, who are uninterested in both topics, should not be a primary target audience, as their diverse interests are harder to specify and align with NutrientH20's brand messaging.

By examining the PCA results, it is evident that the overall data is quite diverse, making it somewhat challenging to clearly differentiate all clusters. However, the analysis still provided valuable insights for advertising strategies. To attract the most relevant audiences, NutrientH20 should focus on launching campaigns that align with health and lifestyle interests. For individuals in Cluster 4 (those with limited engagement in health and lifestyle topics), the company should allocate fewer resources, due to the lack of clear trend of interest.