AeroFit Business Analysis`

Artem Sadreev

2023-02-07

## My Other Projects

Sign Up Here: <https://github.com/ArSadreev/RProjects>

## Executive Summary

The market research team at AeroFit wants to identify the characteristics of the target audience for KP281 treadmill offered by the company, to provide a better recommendation of the treadmills to new customers. The team decides to investigate whether there are differences across the product with respect to customer characteristics. I am perform descriptive analytics to create a customer profile for each AeroFit treadmill product by developing appropriate tables and charts.

Product Portfolio:

The KP281 is an entry-level treadmill that sells for $1,500; The KP481 is for mid-level runners and sells for $1,750; The KP781 treadmill is having advanced features and it sells for $2,500.

## Table of Contents

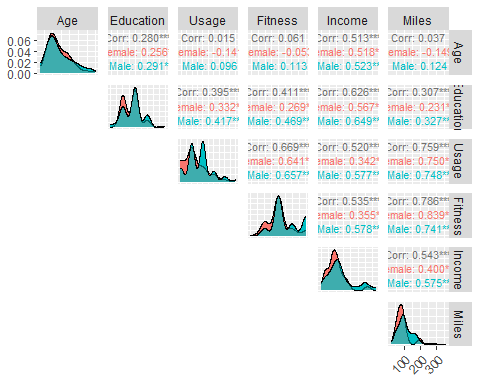
## Data Exploration

Customer population is slightly skewed towards males 57% vs 42% for females.

**Table :** gender distribution

| Gender | population | percentage |
| --- | --- | --- |
| Female | 76 | 42.22222 |
| Male | 104 | 57.77778 |

Figure points out some correlations that might be worth exploring such as the fact that fitness level is correlated with education more for males than females. Also, fitness level and education seems to be more correlated with income for males rather than females. In addition, all variables appear to be skewed to the right for males rather than females.



**Figure** **:** gender correlation map

Table shows us that female users have the highest probability (52.6%) of purchasing KP281 and lowest probability (9%) of purchasing KP781. Since KP781 is the most expensive product selling for 2,500$ and has the most advanced features, the result that we observe can be attributed to lower income distribution of female users compared to male and lower fitness level distribution for female users vs. male users.

**Table** **:** gender distribution by product

| Product | Gender | user\_count | probs |
| --- | --- | --- | --- |
| KP281 | Female | 40 | 0.52631579 |
| KP281 | Male | 40 | 0.38461538 |
| KP481 | Female | 29 | 0.38157895 |
| KP481 | Male | 31 | 0.29807692 |
| KP781 | Female | 7 | 0.09210526 |
| KP781 | Male | 33 | 0.31730769 |

Let’s conduct Chi-squared test to see whether results that we observe are statistically significant. P-value is very close to 0 indicating that our result is statistically significant and gender is indeed related to product choice. So, female users are more likely to buy KP281 and least likely to buy KP781. Male users also have the highest probability to buy KP281 (40%), however probabilities for KP481 and KP781 are not drastically different (29.8% and 31.7% accordingly).

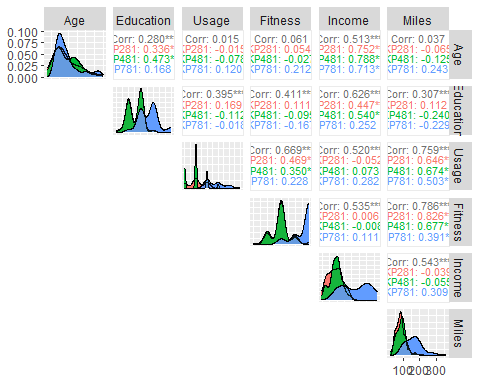
**Table** **:** Chi-squared test

| Gender | KP281 | KP481 | KP781 |
| --- | --- | --- | --- |
| Female | 40 | 29 | 7 |
| Male | 40 | 31 | 33 |

**Table** **:** Chi-squared test

| statistic | p.value | parameter | method |
| --- | --- | --- | --- |
| 12.92384 | 0.001561797 | 2 | Pearson's Chi-squared test |

However, lets explore other variable to see if there is a possible confounding effect and gender is not a primary driver of sales of different products. First of all, education, usage, fitness level, income and miles appear to be skewed to the right for KP781 product indicating that customer base for KP781 have higher level of fitness, higher usage, higher education and higher income when comparing to other products.



**Figure** **:** product correlation map

Let’s calculate probability distribution of customers of different products based on the fitness level. Interestingly, customers with fitness level from 1-3 are more likely to purchase KP281 and KP481 vs. KP781. Customers with fitness level of 4 are roughly equally likely to purchase either product.

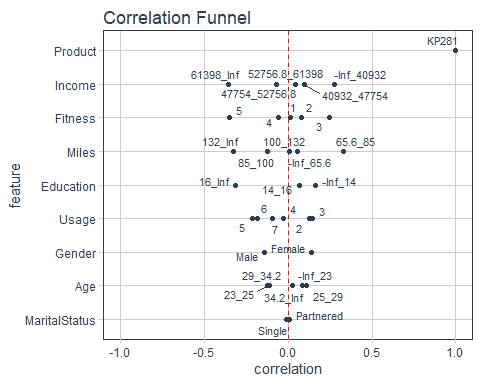
## # A tibble: 12 × 4  
## # Groups: Fitness [5]  
## Product Fitness count prob\_by\_Fitness  
## <chr> <int> <int> <dbl>  
## 1 KP281 1 1 0.5   
## 2 KP281 2 14 0.538   
## 3 KP281 3 54 0.557   
## 4 KP281 4 9 0.375   
## 5 KP281 5 2 0.0645  
## 6 KP481 1 1 0.5   
## 7 KP481 2 12 0.462   
## 8 KP481 3 39 0.402   
## 9 KP481 4 8 0.333   
## 10 KP781 3 4 0.0412  
## 11 KP781 4 7 0.292   
## 12 KP781 5 29 0.935

## Data Modelling for KP281

In this section, I will explore relationships between variables more formally to distinguish core features of customers of different products. In the table below, I am converting all variables to binary format to construct correlation funnel for each product.

## Rows: 10  
## Columns: 36  
## $ Product\_\_KP281 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1  
## $ Product\_\_KP481 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Product\_\_KP781 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ `Age\_\_-Inf\_23` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1  
## $ Age\_\_23\_25 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Age\_\_25\_29 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Age\_\_29\_34.2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Age\_\_34.2\_Inf <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Gender\_\_Female <dbl> 0, 0, 1, 0, 0, 1, 1, 0, 0, 1  
## $ Gender\_\_Male <dbl> 1, 1, 0, 1, 1, 0, 0, 1, 1, 0  
## $ `Education\_\_-Inf\_14` <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 0  
## $ Education\_\_14\_16 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 1  
## $ Education\_\_16\_Inf <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ MaritalStatus\_\_Partnered <dbl> 0, 0, 1, 0, 1, 1, 1, 0, 0, 1  
## $ MaritalStatus\_\_Single <dbl> 1, 1, 0, 1, 0, 0, 0, 1, 1, 0  
## $ Usage\_\_2 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1  
## $ Usage\_\_3 <dbl> 1, 0, 0, 1, 0, 1, 1, 1, 0, 0  
## $ Usage\_\_4 <dbl> 0, 0, 1, 0, 1, 0, 0, 0, 0, 0  
## $ Usage\_\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0  
## $ Usage\_\_6 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Usage\_\_7 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Fitness\_\_1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Fitness\_\_2 <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0  
## $ Fitness\_\_3 <dbl> 0, 1, 1, 1, 0, 1, 1, 1, 0, 1  
## $ Fitness\_\_4 <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0  
## $ Fitness\_\_5 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ `Income\_\_-Inf\_40932` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1  
## $ Income\_\_40932\_47754 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Income\_\_47754\_52756.8 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Income\_\_52756.8\_61398 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Income\_\_61398\_Inf <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ `Miles\_\_-Inf\_65.6` <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0  
## $ Miles\_\_65.6\_85 <dbl> 0, 1, 1, 1, 0, 1, 1, 1, 0, 1  
## $ Miles\_\_85\_100 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Miles\_\_100\_132 <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0  
## $ Miles\_\_132\_Inf <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0

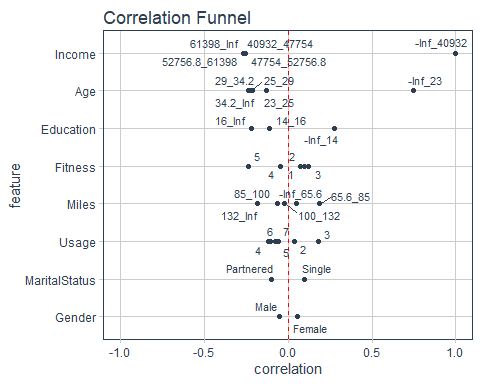
Constructing correlation funnel where the most influencing features for a selected product are shown at the top and least impacting at the bottom. Product KP281 is correlated the most with lower income bracket. So, people who earn below 40,932 are more likely to purchase it. Also, people with fitness level of 3, below 14 years of education, average usage and female of 25-29 years of age are a likely customer for this product. Marital status has little influence in this case.



**Figure** **:** correlation funnel

Since usage and miles basically measure the same, I will drop Miles variable. Also, Marital Status has a very low correlation with KP281 ownership, so I will drop this variable for further analyses as well.

Here we can see that low income bracket is highly correlated with age of 23 and below and education of 14 years and below. However, correlation is below 0.8 for predictors suggesting that we shouldn’t encounter multicollinearity.



**Figure** **:** Income correlation funnel

I am using logistic regression where Age, Fitness and Income variables were separated into 5 different bins, so that we can have a clear picture of what a customer profile for this product might be. Intercept includes customers with 29K to 45K of Income, 18 to 24 years of age and fitness level of 3. I tried out different combination of variables and arrived at only 3 variables that are significant to our results. As it turns out, individuals with 29K to 45K of Income, Fitness level of 3 and 43 to 50 years of age are 10x times more likely to buy KP281 than a baseline customer base. Having income in between 44K and 60K will further increase odds of buying by 0.2x.

**Table** **:** logistic regression

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 2.349769e+00 | 0.4067549 | 2.100324326 | 0.0357003208 |
| Age(24.4,30.8] | 2.469736e+00 | 0.6095233 | 1.483308902 | 0.1379923583 |
| Age(30.8,37.2] | 1.932504e+00 | 0.7270410 | 0.906161412 | 0.3648504242 |
| Age(37.2,43.6] | 9.996865e+00 | 1.0042945 | 2.292426658 | 0.0218810352 |
| Age(43.6,50] | 1.044667e+01 | 1.4741476 | 1.591620504 | 0.1114700002 |
| Fitness1 | 1.631607e+00 | 1.8217779 | 0.268729639 | 0.7881377415 |
| Fitness2 | 7.736329e-01 | 0.5203659 | -0.493225653 | 0.6218531540 |
| Fitness4 | 4.898583e-01 | 0.5919638 | -1.205545127 | 0.2279929279 |
| Fitness5 | 3.068974e-01 | 0.8848669 | -1.334937098 | 0.1818969347 |
| Income(44565.8,59569.6] | 1.996226e-01 | 0.5982969 | -2.693189264 | 0.0070772081 |
| Income(59569.6,74573.4] | 3.520447e-02 | 0.9985120 | -3.351569340 | 0.0008035491 |
| Income(74573.4,89577.2] | 3.579748e-09 | 2396.3913219 | -0.008115525 | 0.9935248190 |
| Income(89577.2,104581] | 2.400059e-09 | 1775.8283410 | -0.011176628 | 0.9910825268 |

Here is a confusion matrix which demonstrates how resulting model performs on my test set.

## Truth  
## Prediction 0 1  
## 0 13 5  
## 1 7 11

And corresponding ROC AUC which is 0.7 indicating that this model has predictive power.

**Table** **:** ROC AUC

| .metric | .estimator | .estimate |
| --- | --- | --- |
| roc\_auc | binary | 0.696875 |

In conclusion, ideal customer profile for KP281 product is someone of fitness level of 3 and 43 to 50 years of age and income in between 44K and 60K. They are 10.6x times more likely to buy than any other customer group.