Research on Factors that affect ski resort pricing

Xinyuan Chen

Yizhan Xue

May 13, 2022

Overview

- ❑ 1. Research Background and Introduction
- ❑ 2. Dataset Description
- ❑ 3. Research Methods Introduction
- ❑ 4. Multiple Linear Regression (MLR)
- ❑ 5. Principal Component Analysis (PCA)
- ❑ 6. Logistic Regression (LR)
- ❑ 7. Result and Conclusion

Research Background and Introduction

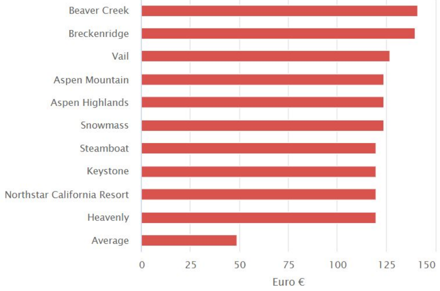
Background:

- Skiing is an increasingly popular sport.
- 6114 ski resorts operated worldwide.
- 470 ski resorts operated in the United States.

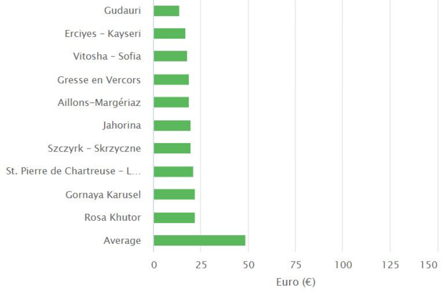
Research Question:

- What factors decide whether a ski resort is an expensive or an economic ski resort?
- What factors do the ski resort managers design their pricing strategy based on?
- How to help tourists choose the most cost-effective ski resorts?
- [Future Work] What factors should investors be most aware of when managing a ski resort?

Most expensive ski resorts



Cheapest ski resorts



Overview

- ❑ 1. Research Background and Introduction
- ❑ **2. Dataset Description**
- ❑ 3. Research Methods Introduction
- ❑ 4. Multiple Linear Regression (MLR)
- ❑ 5. Principal Component Analysis (PCA)
- ❑ 6. Logistic Regression (LR)
- ❑ 7. Conclusion

Dataset Description

- Contains **512** samples and **19** variables.
- Each sample containing information about one ski resort from worldwide.
- DayPassPrice is chosen as the dependent variable, the other 18 variables are independent variables.

Variable	**Definition**	**Data Type**
Ski pass prices adult	The price shows what it costs for 1 adult for 1 day in the main season in Euro €.	int
Beginner slopes	The total amount of "beginner" slopes in kilometer at the resort.	int
Highest point	The highest mountain point at the ski resort	num(number)
…… (19 variables in total)		

Overview

- ❑ 1. Research Background and Introduction
- ❑ 2. Dataset Description
- ❑ **3. Research Methods Introduction**
- ❑ 4. Multiple Linear Regression (MLR)
- ❑ 5. Principal Component Analysis (PCA)
- ❑ 6. Logistic Regression (LR)
- ❑ 7. Conclusion

Research Methods

- Multiple Linear Regression (MLR)
  - The Multiple R-squared is 0.5823
- Principal Component Analysis (PCA)

  Dimensionality Reduction. Show resort facilities and natural location as the two most important dimensions for the research.
- Logistic Regression (LR)
  - The accuracy is 0.88 and the F-1 score is 0.94.
  - DayPrice was changed from a numeric to a nominal variable by using the median as a dividing criterion.

Overview

- ❑ 1. Research Background and Introduction
- ❑ 2. Dataset Description
- ❑ 3. Research Methods Introduction
- ❑ **4. Multiple Linear Regression (MLR)**
- ❑ 5. Principal Component Analysis (PCA)
- ❑ 6. Logistic Regression (LR)
- ❑ 7. Conclusion

Method 1: Multiple Linear Regression (MLR)

(1) Checking the Assumptions

- Linear Relationship
- Multivariate Normality
- Multicollinearity
- Homoscedasticity

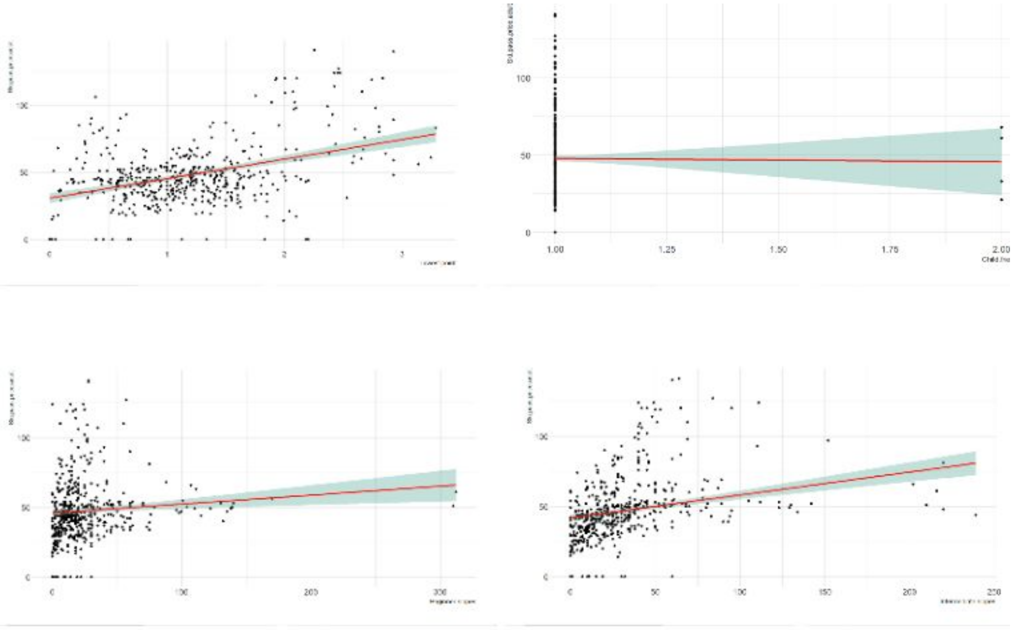
- Draw scatter plots

  Remove variables which has non-linear relationship with target value or use stepwise regression.
- Examine a normal Predicted Probability (P-P) plot
- Checking the VIF

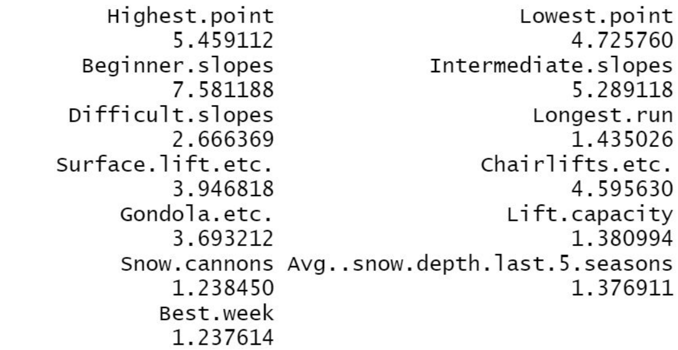
  Make sure VIF < 5 for all independent variables to ensure no multicollinearity.
- Examine the scatterplot of the residuals

Method 1: Multiple Linear Regression (MLR)

- Scatterplots between independent variables and dependent variable

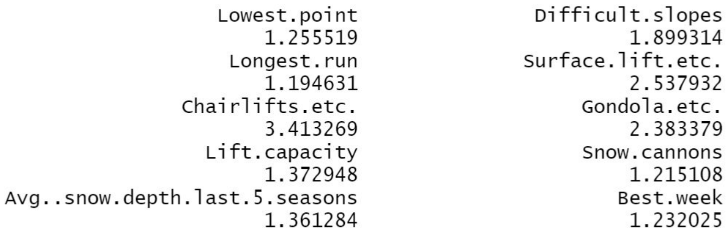


- Checking VIF



```
          Highest.point                     Lowest.point
               5.459112                         4.725760
       Beginner.slopes              Intermediate.slopes
               7.581188                         5.289118
       Difficult.slopes                      Longest.run
               2.666369                         1.435026
     Surface.lift.etc.                 Chairlifts.etc.
               3.946818                         4.595630
          Gondola.etc.                    Lift.capacity
               3.693212                         1.380994
          Snow.cannons Avg..snow.depth.last.5.seasons
               1.238450                         1.376911
              Best.week
               1.237614
```

Before Removal

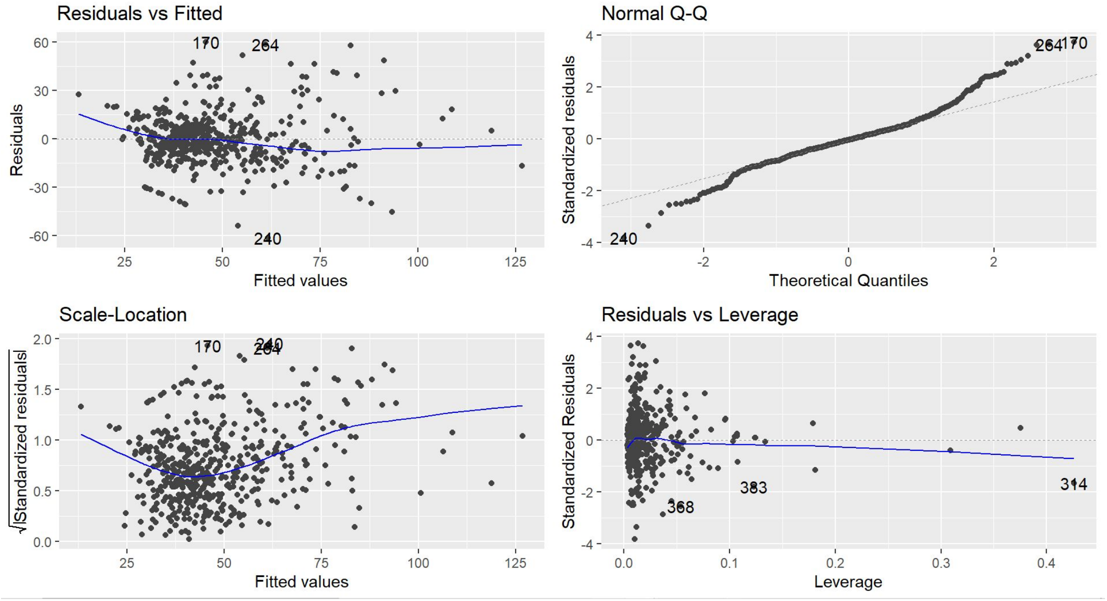


```
                   Lowest.point         Difficult.slopes
                       1.255519                 1.899314
                    Longest.run        Surface.lift.etc.
                       1.194631                 2.537932
                Chairlifts.etc.             Gondola.etc.
                       3.413269                 2.383379
                  Lift.capacity             Snow.cannons
                       1.372948                 1.215108
Avg..snow.depth.last.5.seasons                Best.week
                       1.361284                 1.232025
```

After Removal

Method 1: Multiple Linear Regression (MLR)

- Diagnostic Plots

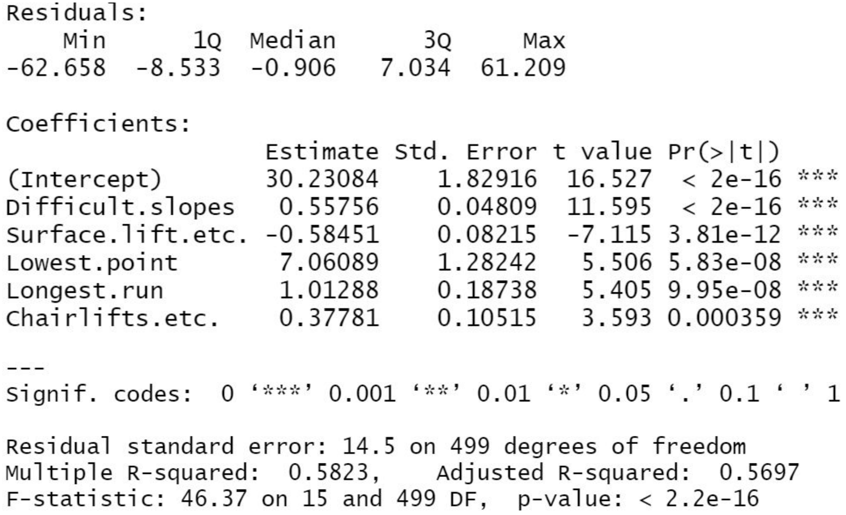


- The dataset is not totally satisfied with the assumption of the MLR except no multicollinearity.

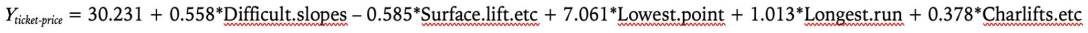
↓

- Use stepwise regression to select variables for better performance. Number of independent variables from 15 to 5.

Method 1: Multiple Linear Regression (MLR)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-62.658  -8.533  -0.906   7.034  61.209 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       30.23084    1.82916  16.527  < 2e-16 ***
Difficult.slopes   0.55756    0.04809  11.595  < 2e-16 ***
Surface.lift.etc. -0.58451    0.08215  -7.115 3.81e-12 ***
Lowest.point       7.06089    1.28242   5.506 5.83e-08 ***
Longest.run        1.01288    0.18738   5.405 9.95e-08 ***
Chairlifts.etc.    0.37781    0.10515   3.593 0.000359 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.5 on 499 degrees of freedom
Multiple R-squared:  0.5823,	Adjusted R-squared:  0.5697 
F-statistic: 46.37 on 15 and 499 DF,  p-value: < 2.2e-16
```



$$Y_{ticket\text{-}price} = 30.231 + 0.558*\text{Difficult.slopes} - 0.585*\text{Surface.lift.etc} + 7.061*\text{Lowest.point} + 1.013*\text{Longest.run} + 0.378*\text{Charlifts.etc}$$

- The linear model is statistically significant with 0.05 level.
  - Check the p-value.
- The dependent variable has not sufficiently been explained by this model.
  - The R-squared < 0.7.
  - May be because it is not satisfied all assumptions.
- Can be used as a baseline model to suggest a reliable daily ticket price.

Overview

❑ 1. Research Background and Introduction

❑ 2. Dataset Description

❑ 3. Research Methods Introduction

❑ 4. Multiple Linear Regression (MLR)

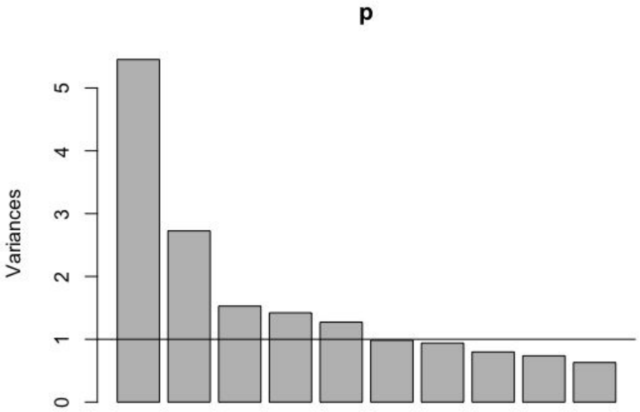
❑ **5. Principal Component Analysis (PCA)**

❑ 6. Logistic Regression (LR)

❑ 7. Conclusion

Method 2: Principal Component Analysis (PCA)

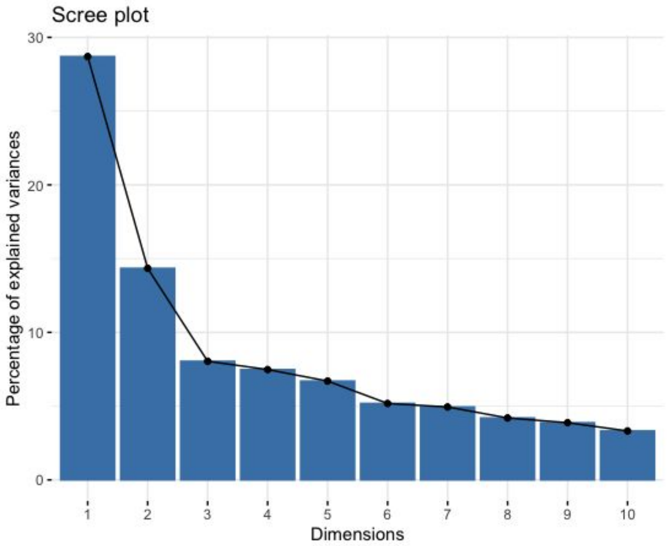
Step 1: Determine the Number of Components



The Kaiser-Meyer-Olkin Method

Method 2: Principal Component Analysis (PCA)

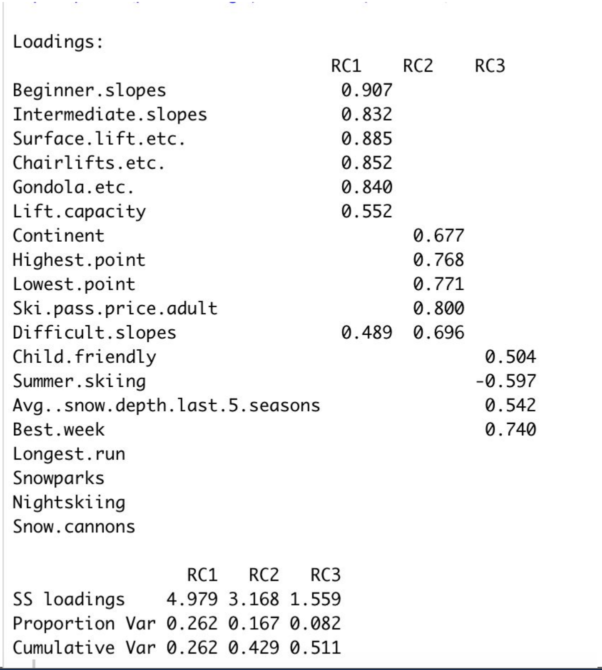
Step 1: Determine the Number of Components (cont)



The Knee Method

Method 2: Principal Component Analysis (PCA)

Step 2: Determine the Rotate Components



```
Loadings:
                              RC1    RC2    RC3   
Beginner.slopes                0.907              
Intermediate.slopes            0.832              
Surface.lift.etc.              0.885              
Chairlifts.etc.                0.852              
Gondola.etc.                   0.840              
Lift.capacity                  0.552              
Continent                             0.677       
Highest.point                         0.768       
Lowest.point                          0.771       
Ski.pass.price.adult                  0.800       
Difficult.slopes               0.489  0.696       
Child.friendly                               0.504
Summer.skiing                               -0.597
Avg..snow.depth.last.5.seasons               0.542
Best.week                                    0.740
Longest.run                                       
Snowparks                                         
Nightskiing                                       
Snow.cannons                                      

                 RC1   RC2   RC3
SS loadings    4.979 3.168 1.559
Proportion Var 0.262 0.167 0.082
Cumulative Var 0.262 0.429 0.511
```

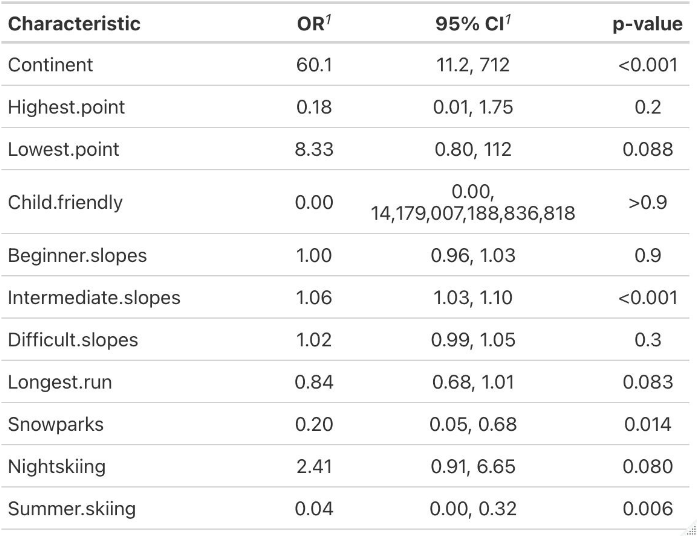
$Y_{\text{tourist-feeling}}$ = 0.907*Beginner.slopes +
0.832*Intermediate.slopes +
0.885*Surface.lift.etc. + 0.852*Chairlifts.etc. +
0.840*Gondola.etc. + 0.552*Lift.capacity +
0.489*Difficult.slopes

Overview

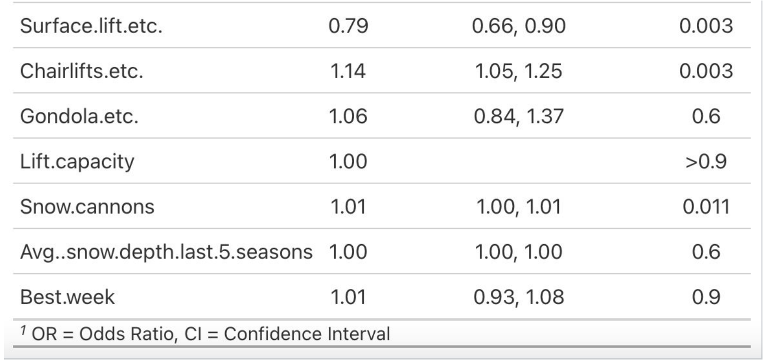
- □ 1. Research Background and Introduction
- □ 2. Dataset Description
- □ 3. Research Methods Introduction
- □ 4. Multiple Linear Regression (MLR)
- □ 5. Principal Component Analysis (PCA)
- □ **6. Logistic Regression (LR)**
- □ 7. Conclusion

Method 3: Logistic Regression (LR)

Step 1: Output the Odds Ratio Indexes



Characteristic	OR[1]	95% CI[1]	p-value
Continent	60.1	11.2, 712	<0.001
Highest.point	0.18	0.01, 1.75	0.2
Lowest.point	8.33	0.80, 112	0.088
Child.friendly	0.00	0.00, 14,179,007,188,836,818	>0.9
Beginner.slopes	1.00	0.96, 1.03	0.9
Intermediate.slopes	1.06	1.03, 1.10	<0.001
Difficult.slopes	1.02	0.99, 1.05	0.3
Longest.run	0.84	0.68, 1.01	0.083
Snowparks	0.20	0.05, 0.68	0.014
Nightskiing	2.41	0.91, 6.65	0.080
Summer.skiing	0.04	0.00, 0.32	0.006

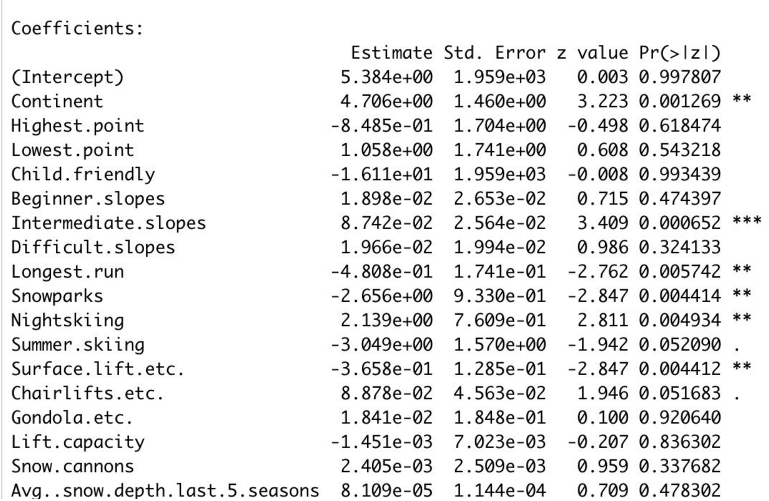


Surface.lift.etc.	0.79	0.66, 0.90	0.003
Chairlifts.etc.	1.14	1.05, 1.25	0.003
Gondola.etc.	1.06	0.84, 1.37	0.6
Lift.capacity	1.00		>0.9
Snow.cannons	1.01	1.00, 1.01	0.011
Avg..snow.depth.last.5.seasons	1.00	1.00, 1.00	0.6
Best.week	1.01	0.93, 1.08	0.9

[1] OR = Odds Ratio, CI = Confidence Interval

Method 3: Logistic Regression (LR)

Step 1: Output the Odds Ratio Indexes (cont)

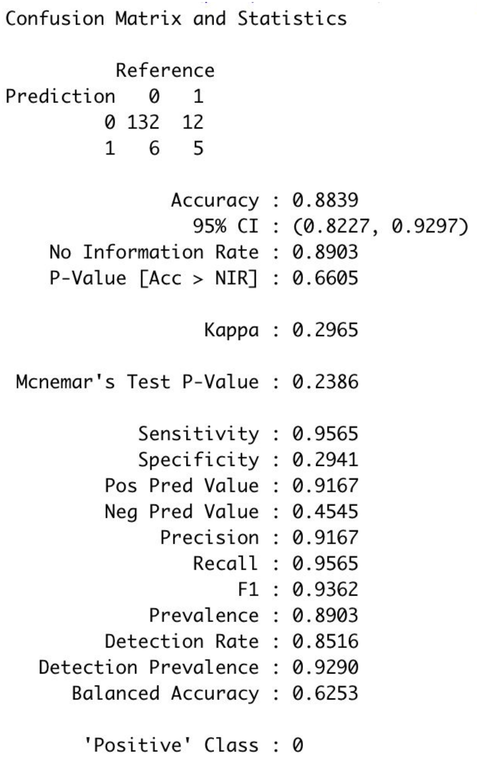


```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   5.384e+00  1.959e+03   0.003 0.997807    
Continent                     4.706e+00  1.460e+00   3.223 0.001269 ** 
Highest.point                -8.485e-01  1.704e+00  -0.498 0.618474    
Lowest.point                  1.058e+00  1.741e+00   0.608 0.543218    
Child.friendly               -1.611e+01  1.959e+03  -0.008 0.993439    
Beginner.slopes               1.898e-02  2.653e-02   0.715 0.474397    
Intermediate.slopes           8.742e-02  2.564e-02   3.409 0.000652 ***
Difficult.slopes              1.966e-02  1.994e-02   0.986 0.324133    
Longest.run                  -4.808e-01  1.741e-01  -2.762 0.005742 ** 
Snowparks                    -2.656e+00  9.330e-01  -2.847 0.004414 ** 
Nightskiing                   2.139e+00  7.609e-01   2.811 0.004934 ** 
Summer.skiing                -3.049e+00  1.570e+00  -1.942 0.052090 .  
Surface.lift.etc.            -3.658e-01  1.285e-01  -2.847 0.004412 ** 
Chairlifts.etc.               8.878e-02  4.563e-02   1.946 0.051683 .  
Gondola.etc.                  1.841e-02  1.848e-01   0.100 0.920640    
Lift.capacity                -1.451e-03  7.023e-03  -0.207 0.836302    
Snow.cannons                  2.405e-03  2.509e-03   0.959 0.337682    
Avg..snow.depth.last.5.seasons  8.109e-05  1.144e-04   0.709 0.478302    
```

$$P = \frac{e^{(5.394e+00)+(4.706e+00)*Continent+(8.742e-02)*Intermediate.slopes-(4.808e+01)*Longest.run-(2.656e-00)*Snoparks+(2.139e+00)*Nightskiing-(3.658e+01)*Surface.lift.etc}}{1+\ e^{(5.394e+00)+(4.706e+00)*Continent+(8.742e-02)*Intermediate.slopes-(4.808e+01)*Longest.run-(2.656e-00)*Snoparks+(2.139e+00)*Nightskiing-(3.658e+01)*Surface.lift.etc}}$$

Method 3: Logistic Regression (LR)

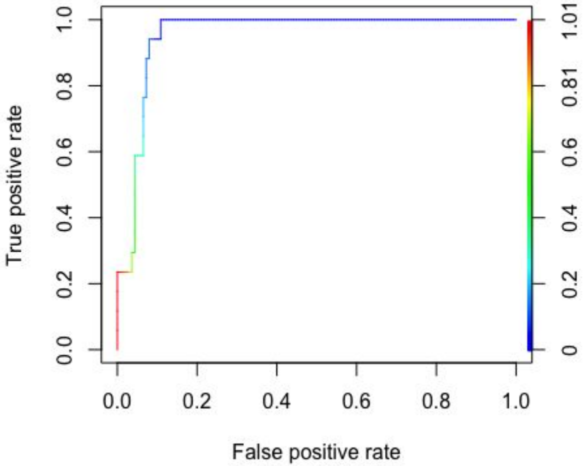
Step 2: Further Predicting Through Classification



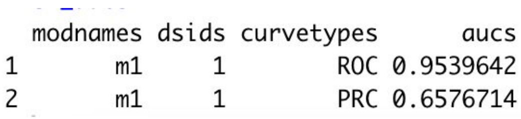
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 132  12
         1   6   5

               Accuracy : 0.8839
                 95% CI : (0.8227, 0.9297)
    No Information Rate : 0.8903
    P-Value [Acc > NIR] : 0.6605

                  Kappa : 0.2965

 Mcnemar's Test P-Value : 0.2386

            Sensitivity : 0.9565
            Specificity : 0.2941
         Pos Pred Value : 0.9167
         Neg Pred Value : 0.4545
              Precision : 0.9167
                 Recall : 0.9565
                     F1 : 0.9362
             Prevalence : 0.8903
         Detection Rate : 0.8516
   Detection Prevalence : 0.9290
      Balanced Accuracy : 0.6253

       'Positive' Class : 0
```

Method 3: Logistic Regression (LR)

Step 3: Output the ROC Curve and AUC values



ROC Curve



```
  modnames dsids curvetypes      aucs
1       m1     1        ROC 0.9539642
2       m1     1        PRC 0.6576714
```

AUC values

Comparison of the three models

Multiple Linear Regression

Advantages:

The ability to determine the relative influence of one or more predictor variables to the criterion value.

Easy to identify outliers, or anomalies.

Disadvantages:

Outliers can have huge effects on the linear regression model;

Looks at a relationship between the mean of the dependent variables and the independent variables.

Principal Component Analysis

Advantages:

Easy to compute;

Speeds up other machine learning algorithms;

Counteracts the issues of high-dimensional data.

Disadvantages:

Low interpretability of principal components;

The trade-off between information loss and dimensionality reduction.

Logistic Regression

Advantages:

Performs well when the dataset is linearly separable;

Less prone to overfitting but it can overfit in high dimensional datasets;

Disadvantages:

Can only be used to predict discrete functions;

Overview

❑ 1. Research Background and Introduction

❑ 2. Dataset Description

❑ 3. Research Methods Introduction

❑ 4. Multiple Linear Regression (MLR)

❑ 5. Principal Component Analysis (PCA)

❑ 6. Logistic Regression (LR)

❑ **7. Conclusion**

Conclusion

Limitation:
Dataset Limitation;
The characteristics of factors are not obvious enough;
Too much missing data;

Suggestions and Future Work:
Safety factors are most important to consider;
Choose a place with a relatively low altitude to build a ski resort;
Various emergency measures to protect safety need to be followed up

Conclusion:
Resorts' facilities, and resorts location are the factors that tourists concern most when choosing a ski resort as a destination.
The more advanced the ski facilities, and the better the location of the resort, the higher the ticket price will be.

Reference

Berard-Chenu, L., Cognard, J., François, H., Morin, S., & George, E. (2021). *Do changes in snow conditions have an impact on snowmaking investments in French Alps ski resorts?* International Journal of Biometeorology, 65(5), 659-675.

Boris, D., Makajić-Nikolić, D., Ćirović, M., Petrović, N., & Suknović, M. (2020). *A ski injury risk assessment model for ski resorts. Journal of Risk Research*, 23(12), 1590–1602. https://doi-org.proxy2.library.illinois.edu/10.1080/13669877.2020.1749113

Carus, L., & Castillo, I. (2021). *Managing risk in ski resorts: Environmental factors affecting actual and estimated speed on signposted groomed slopes in a cohort of adult recreational alpine skiers*. Plos one, 16(8), e0256349.

Eberly, L. E. (2007). Multiple linear regression. *Topics in Biostatistics*, 165-187.

Hatzenbuehler, J., Glazer, J., & Kuhn, C. (2009). *Awareness of Altitude Sickness Among Visitors to a North American Ski Resort*. Wilderness & Environmental Medicine, 20(3), 257–260. https://doi-org.proxy2.library.illinois.edu/10.1580/08-WEME-OR-191R.1

Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: a review and recent developments*. Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 374(2065), 1–16.

Lange, D. (2020, November 25). *Number of ski resorts United States 1990-2020*. Statista. Retrieved May 12, 2022, from https://www.statista.com/statistics/206534/number-of-ski-resorts-operating-in-the-us-since-1990/

Ormiston, D., Gilbert, A., & Manning, R. E. (1998). *Indicators and standards of quality for ski resort management*. *Journal of travel research*, *36*(3), 35-41.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. The Journal of Educational Research, 96(1), 3–14.

Ski-resort-stats. (2020, April 12). *Find resort*. Ski resort statistics. Retrieved May 5, 2022, from https://ski-resort-stats.com/find-ski-resort/

Vanat , L. (2021, April). *2021 International Report on Snow &amp; Mountain Tourism*. vanat.ch. Retrieved May 11, 2022, from https://www.vanat.ch/RM-world-report-2021.pdf.

Thank You
Q & A

Method 2: Principal Component Analysis (PCA)

Advantages:

1) Easy to compute;
2) Speeds up other machine learning algorithms;
3) Counteracts the issues of high-dimensional data.

Disadvantages:

1) Low interpretability of principal components;
2) The trade-off between information loss and dimensionality reduction.

Logistic Regression (LR)

Advantages:

1) Logistic Regression performs well when the dataset is linearly separable;
2) Logistic regression is less prone to overfitting but it can overfit in high dimensional datasets.
3) Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative)

Disadvantages:

1) If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to overfit.
2) Logistic Regression can only be used to predict discrete functions.

Method 2: Principal Component Analysis (PCA)

What does PCA mean?

One of the most important dimensionality-reduction methods. Its main principle is to find the most prominent dimensions in the data and replace the original data with the most important aspects of the data.