Predicting Climbing Success on Himalayan Expeditions

Springboard – Capstone Project 1

Jacques Poolman

September 2019

Table of contents

List of tables

List of figures

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

This is the first project of two capstone projects that forms part of the Data Science Career Track course offered by Springboard. This project aims to showcase the skills learnt during the course thus far by answering interesting real-life data science questions.

In considering various datasets, one question remained seemingly interesting. Many recent news articles covered the overcrowded trails on Mount Everest that led to disastrous consequences for many trying to summit the highest mountain in the world. Of course, Mount Everest is not the only trek in the Himalayas and not all unsuccessful summit attempts lead to your demise. It is still interesting, however, to analyse what contributes to successful summit attempts in Himalayan mountain expeditions. Other than grit and sheer determination, what will give you the best possible chance of successfully summiting a Himalayan mountain trek of your choice?

1.2 PROBLEM STATEMENT

This project attempts to highlight key features that have contributed to successful summit attempts in the past and use that data to aid prospective climbers and enthusiasts in predicting the success of future summit attempts.

According to the website “The Himalayan Database ©” http://himalayandatabase.com/index.html, a website that tracks all Himalayan expeditions since 1905, and the source of the datasets for this project: “The records in the Himalayan Database will be of considerable significance to climbers planning expeditions, to journalists and mountaineering historians needing ready access to historical records, and to medical researchers elucidating patterns of accidents, fatalities, and supplemental oxygen use.”

CHAPTER 2
DATASETS

2.1 DATA COLLECTION

The Himalayan Database ©, a non-profit organisation that was established to continue the work of Elizabeth Hawley, maintains a database comprising all expeditions from 1905 to 2018. This database includes over 450 Nepalese peaks, including Everest, Cho Oyu and Makalu. Each peak contains information particular to the peak itself, such as location and height. Each expedition, in turn, contains information about the expedition, which peak it attempted to summit and information about each individual member. A member's information includes biographical information like age, country of origin, oxygen use and most importantly, summit success.

To access this data it is necessary to download The Himal Program from The Himalayan Database website and select the required data tables for download in CSV format. Figure 2.1 describes the relationship of the data tables.

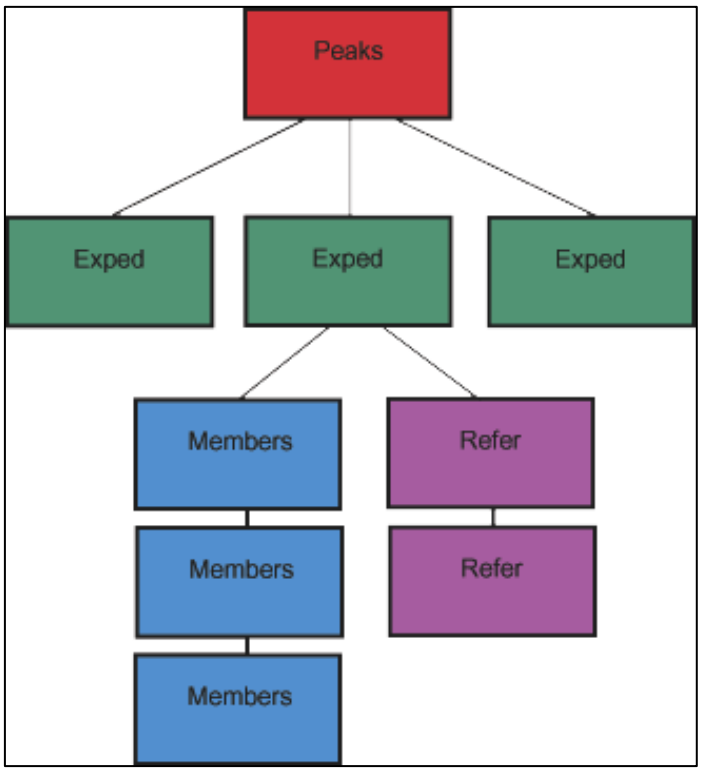

Figure 2.1: Data tables relationship

Source: The Himalayan Database, 2017; Richard Salisbury.

The data tables required for the analysis are 'Peaks', 'Exped' (expeditions) and 'Members'. The table 'Refer' describes the literature references for each expedition and is not required.

2.2 DATA WRANGLING

Appendix A (Tables A.2, A.4 and A.6) presents all three data tables with a description of each feature. The final dataset (Table A.1) includes 51 features from the original three data tables. The rest of the features are dropped.

The output/labelled data is the feature ‘msuccess’, contained in the Members table. This feature describes the success (true or false) of each member that attempted to summit and negates other features relating to the outcome of successful summits.

Some features are already categorised in the ‘The Himalayan Database’ guide and are included in Appendix A, Tables A.3, A.5 and A.7.

2.2.1 Peaks (peaks.csv)

This table contains 468 records, one for each peak, and 22 features that describe the mountaineering peaks of Nepal (Table A.2). Only eight relevant features are included.

The included features describe the geographical nature and status of the peaks which are cleaned and converted to categorical data types, where required.

2.2.2 Exped (expeditions.csv)

This table describes each of the 9,959 climbing expeditions with 65 features each (Table A.4). The table relates to the peaks table through the feature ‘peakid’ and with the members table through the feature ‘membid’. Only 21 relevant features are included.

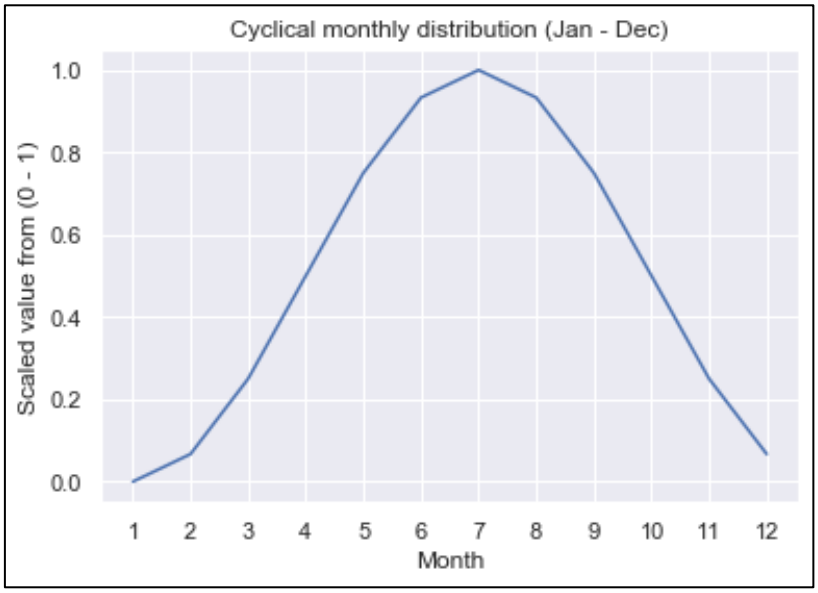

Figure 2.2: Cyclical adaption to month of the year

Source: Author, 2019.

These features describe the makeup of each expedition as well as the when, where and how the expeditions took place. Due to the cyclical nature of ‘month’ and ‘season’, simple categorisation is not enough. To compensate, the data is converted by a sine function to maintain the cyclical nature of the data. Figure 2.2 shows how this is done for the ‘month’ feature.

Some features contains too many NaN values and are dropped. The rest of the features are mostly Boolean type data (Yes/No) and is converted to binary.

2.2.3 Members (members.csv)

This table has 85 features and describes each one of the 65,534 members from all the expeditions (Table A.6). This data contains the who and how for each member, mostly described by biographical information and individual data as captured in the Exped table. For example, an expedition takes oxygen supply with, which could be important for successful summitting, but not all members necessarily use it. Perhaps some members only use oxygen in the climb. Members also attempt to summit by different means, some have more help (Sherpas) than others and some attempts traversing, skiing or gliding down from the summit. How someone descends does not necessarily explain how a member successfully summits, but it might capture some personality traits that could be interesting.

2.2.4 Final pre-processing

After dropping the unwanted features, the three data tables are merged into one dataset by their respective field ID’s, ready for further cleaning.

The data does contain some NaN values that needs adapting. Where data is of binary type, the NaN values are simply imputed with the average value if it does not significantly alter the mean value. Other features are more difficult. The NaN values for a member’s age requires imputing a distribution of mean values over the series to maintain the smooth normal distribution. Where categorical NaN values contributes less than 1% of the data, the whole rows are dropped.

Many features contain too many non-unique values. Unfortunately features such as occupation and agency (the company used for the expedition) cannot be sufficiently categorised without losing a significant amount of data. However, some categories in features like nation, citizen and residence can be reduced to 50 categories with a data loss of 1%, or less.

Finally, all categorical data types are transformed with dummy variables to have a final shape of 63,113 entries with 178 features.

CHAPTER 3
EXPLORATORY DATA ANALYSIS

3.1 ANALYSIS

The final wrangled dataset contains mostly binary data types but there is enough continuous and categorical data available to gain more insight.

Himalayan peaks are spread across Nepal, China and India. Figure 3.1 shows that most peaks are in Nepal. Although Nepal and China share the most peaks together, every single peak extends into Nepal. On the other hand, the region with the most peaks is the Khumbu-Rolwaling-Makalu region (Figure 3.2).

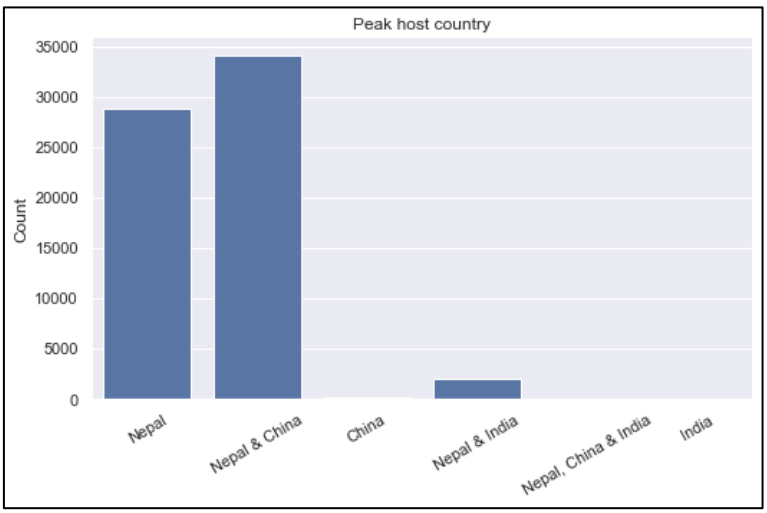

Figure 3.1: Peak Host Country

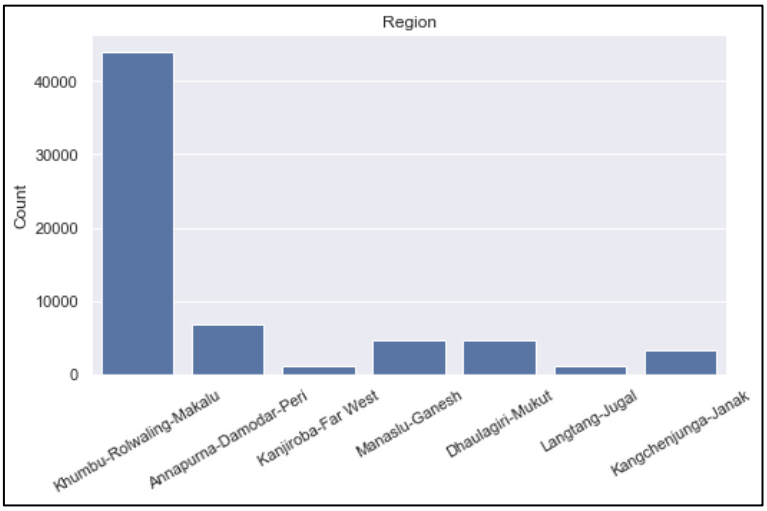

Figure 3.2: Region

Most expeditions are from the USA, Japan and Europe (Figure 3.3) and climbers have a mean age of 36 years old (Figure 3.4).

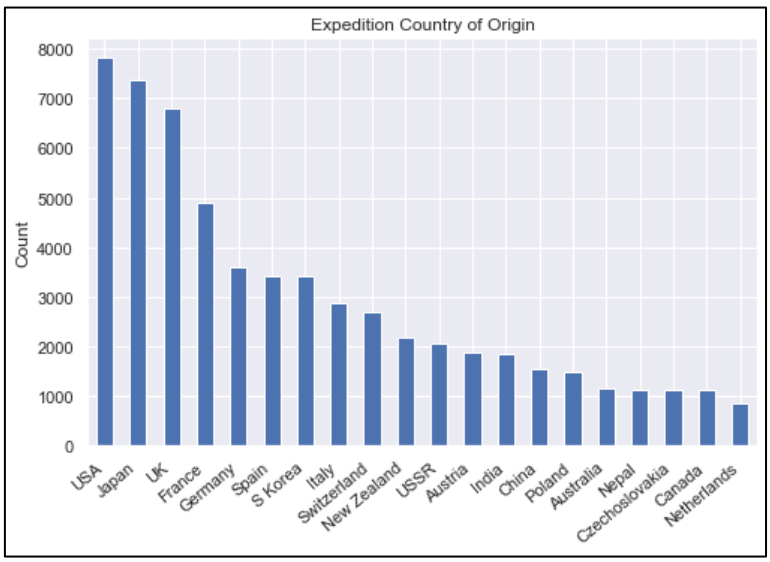

Figure 3.3: Expedition Country of Origin

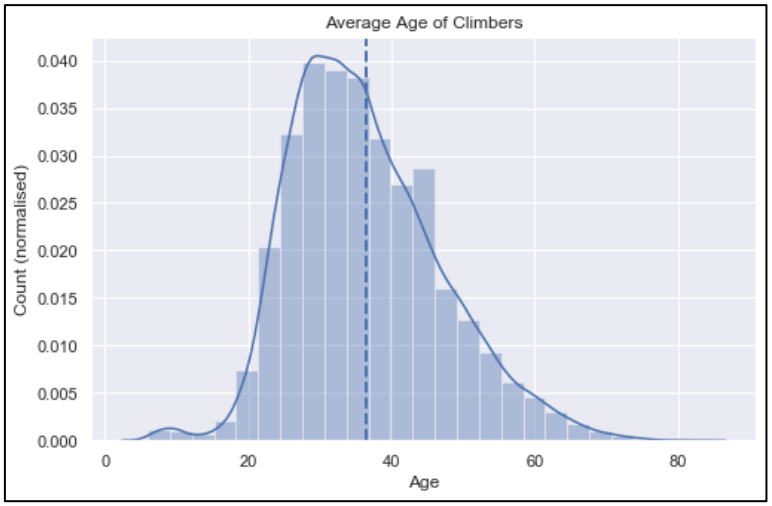

Figure 3.4: Mean Age of Climbers

Figure 3.5 shows how the age range increased over the years and more women started joining expeditions. Interesting are the gaps in the data showing the years with no expeditions. This could be explained by World War 1, the Great Depression, World War 2 and perhaps the Vietnam War.

Reviewing the labelled data of successful member summits, Figure 3.6 offers greater insight into successful summits breakdown between expeditions and individual members. Clearly, not all members in an expedition are successful in reaching the summit even though the expedition is successful. In fact, of all successful expeditions, 16,562 members were unsuccessful. Interestingly, of all unsuccessful expeditions, 9 members were successful in reaching the summit.

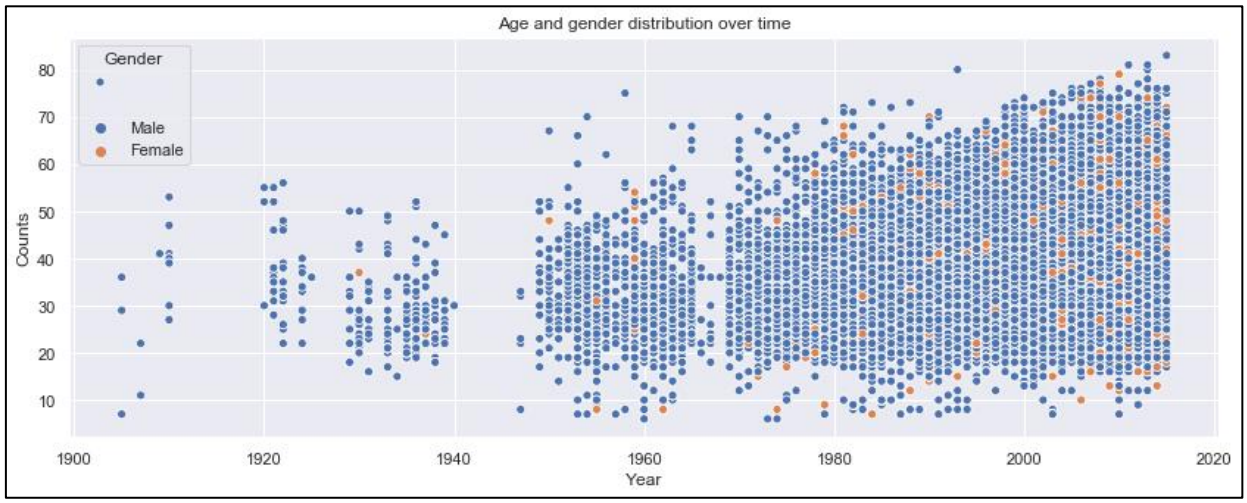

Figure 3.5: Age and Gender over Time

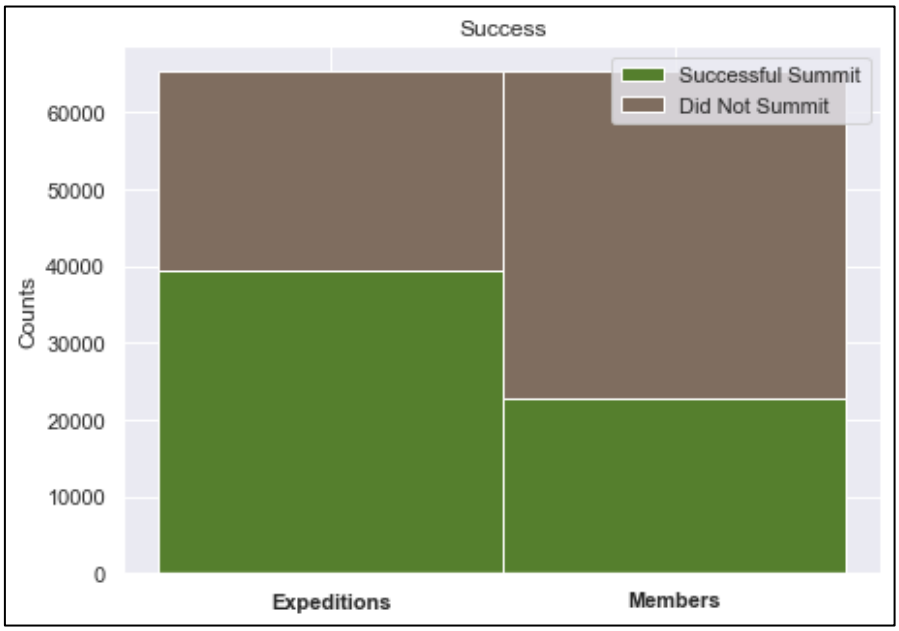

Figure 3.6: Age and Gender over Time

3.2 SUPPORTING STATISTICS

To analyse the data that best describes what contributes to summit success, it is vital to investigate how some features influence this success. Correlations between features give some insight and is used to identify features for further analysis. Hypothesis testing answers some questions and highlights significant differences in some features.

3.2.1 Explore correlations

The correlation matrix in Figure 3.7 graphically presents the Pearson Coefficients between some features and clearly highlights how these features correlate. The labelled data feature of interest is ‘msuccess’.

Comparing this feature on the x-axis with other features on the y-axis seems to indicate some positive correlations with the following: (note, not all features are included)

- ‘stdrte’: If the route to the summit was classified as the standard route. (Perhaps look at other route classifications as well)
- ‘year’: The year the expedition took place.
- ‘sherpa’: If the member was a Sherpa. (Look at hired guides as well)
- General oxygen use including used during climbing and sleeping.

There seems to be some negative, although not significant, correlation between ‘o2none’, meaning no oxygen used.

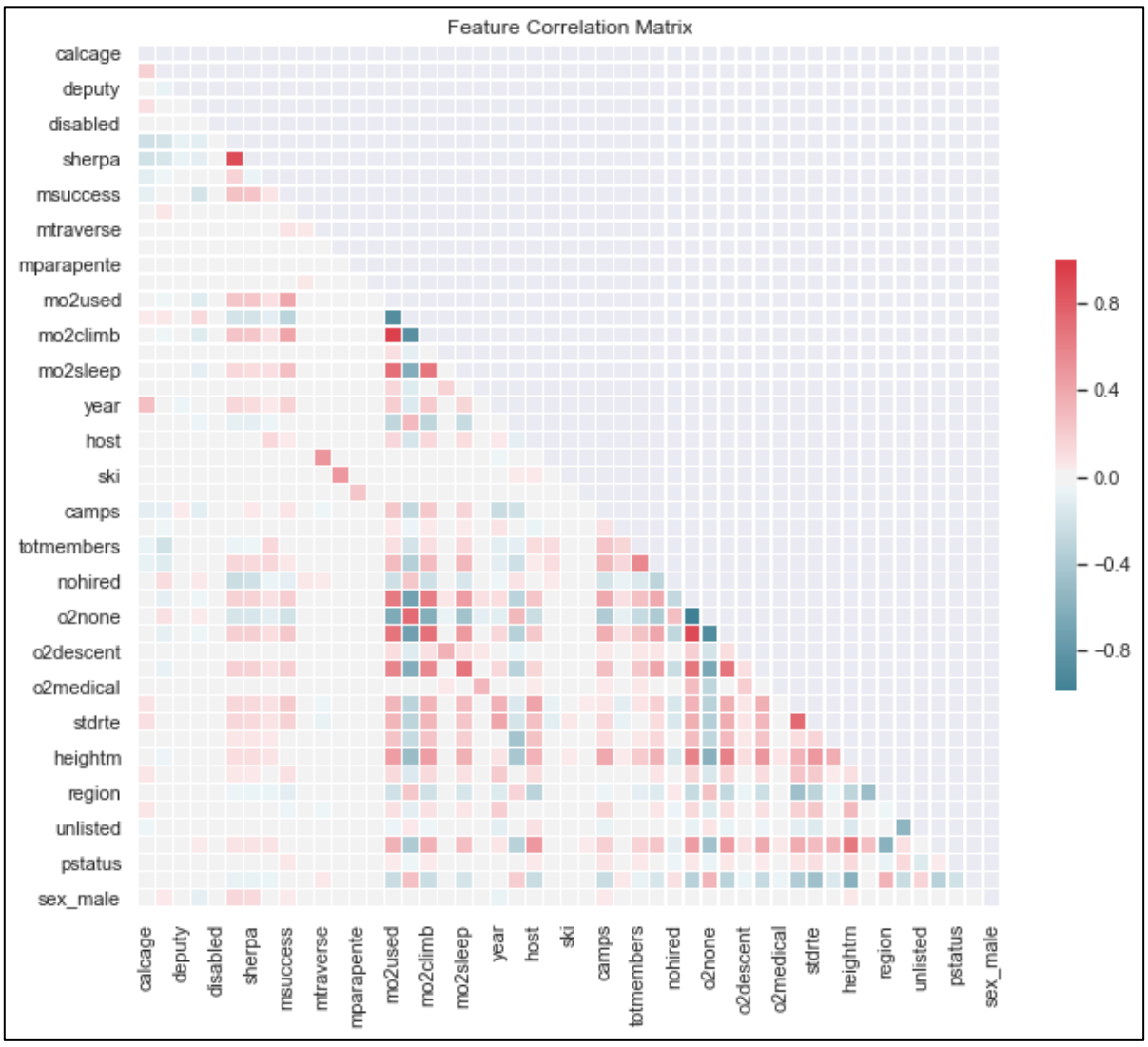

Figure 3.7: Correlation Matrix for Features

3.2.2 Hypothesis Tests

Using some of the features identified by the correlation matrix, among others, the question of significance is formally answered by applying hypothesis testing on the effect these features have on successful outcomes: A member's age, expedition size, number of guides, historic year of summitting and if the route was a commercial route.

All hypothesis test assumed equal variances.

i. Age

The null hypothesis assumes there is no difference between the mean age of members that successfully summited and members that didn't.

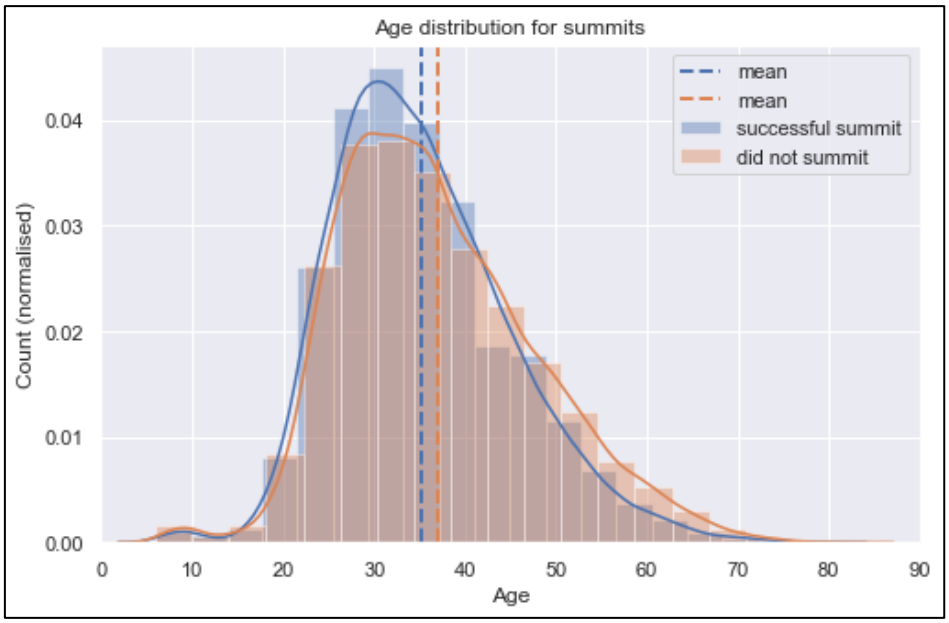

Figure 3.8: Significance of Age on Successful Summits

Figure 3.8 does show a difference in the mean age. Successful members have a mean age of 35.26 and unsuccessful members 37.02. The results from a t-test revealed a p-value of 5.64e-93, or nearly zero. Therefore, reject the null hypothesis; age has a significant effect on the successful outcome of summit attempts. Younger members are more likely to summit.

ii. Expedition size

The null hypothesis assumes there is no difference of the mean size of expeditions (the number of members in an expedition) between successful and unsuccessful summits.

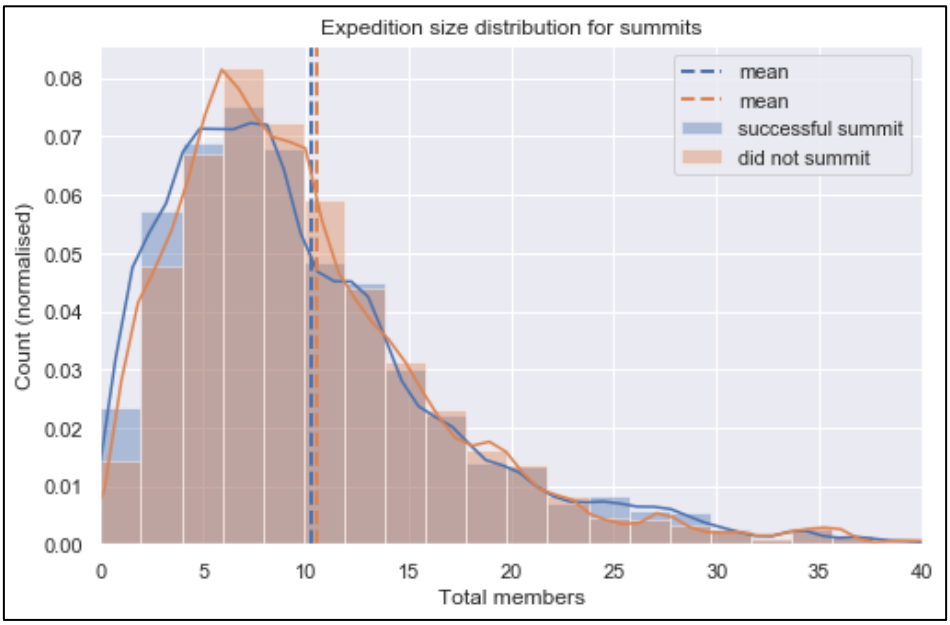

Figure 3.9: Significance of Expedition Size on Successful Summits

Figure 3.9 shows a slight difference in mean expedition size. Successful summits have a mean expedition size of 10.23 members and unsuccessful summits 10.52 members per expedition. The results from a t-test revealed a p-value of 6.63e-05, or nearly zero. Therefore, reject the null hypothesis; expedition size has a significant effect on the successful outcome of summit attempts. Smaller expeditions are more likely to summit.

iii. Number of guides

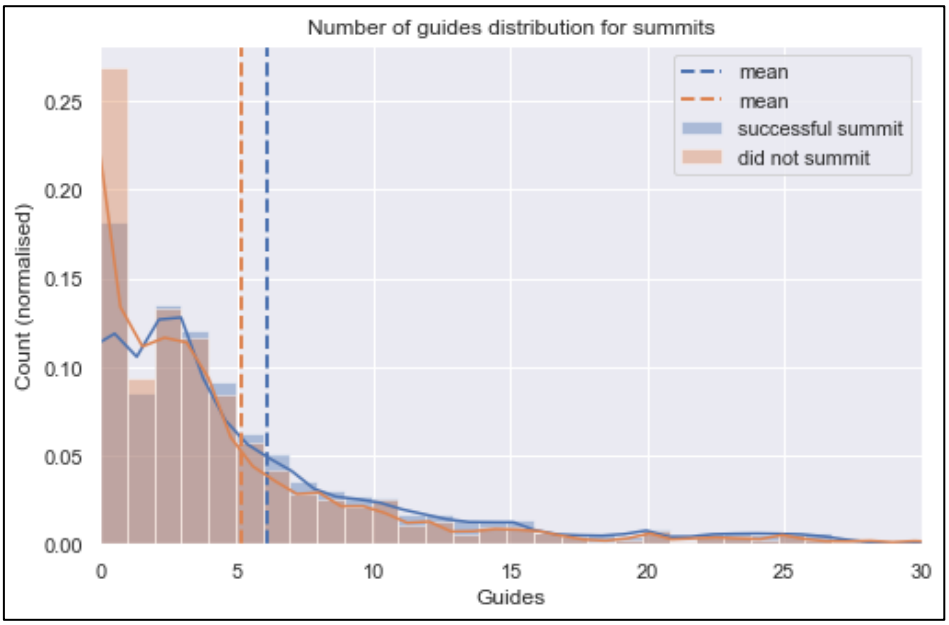

Figure 3.10: Significance of Number of Guides on Successful Summits

The null hypothesis assumes there is no difference of the mean number of guides hired per expedition, between successful and unsuccessful summits.

Figure 3.10 indicates a difference in the mean number of guides hired. Successful summits have a mean number of guides hired of 6.09 per expedition and unsuccessful summits have 5.14 guides. The results from a t-test revealed a p-value of 7.30e-36, or nearly zero. Therefore, reject the null hypothesis; the number of guides hired has a significant effect on the successful outcome of summit attempts. Expeditions with more hired guides are more likely to summit.

iv. Historic approach

The historic approach investigates if summit success has changed over the years. Were you more likely to summit in the past, or in more recent years? Although this question does not necessarily contribute to the probability of immediate success, it is interesting to know if it has become 'easier' to summit. The null hypothesis therefore assumes there is no difference in the mean year, between successful and unsuccessful summits.

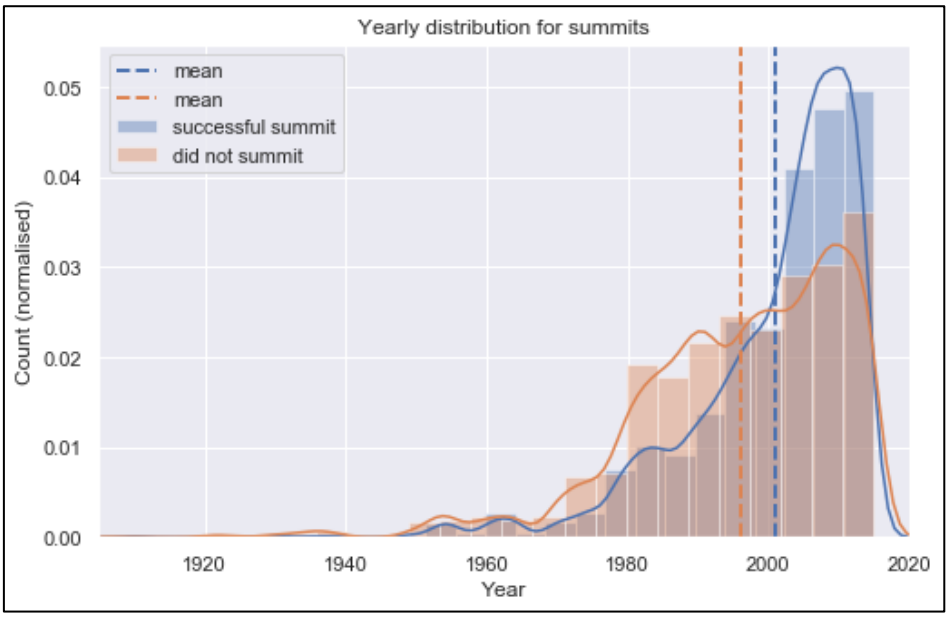

Figure 3.11: Historic Approach on Successful Summits

Figure 3.11 indicates a difference in the mean year. Successful summits have a mean year of 2001 and unsuccessful summits a mean year of 1996. The results from a t-test revealed a p-value of 0.0. Therefore, reject the null hypothesis; when, or how recent a member attempted a summit has a significant effect on success. Members attempting to summit today have a better chance of success compared to attempts in the past.

v. Commercial routes

Commercial routes describe the status of a trekking route as a recognised commercial route, or not. If this binary data type is approached probabilistically, a hypothesis test can determine significance between the mean probability of successful summits occurring on commercial routes. The null hypothesis then assumes there is no difference between the mean probability of success on commercial and non-commercial routes.

Analysis, however, does reveal a difference in the mean probability. Successful summits have a mean probability that a summit attempt occurred on a commercial route as 0.59 and only 0.37 probability for unsuccessful summit attempts. The results from a t-test revealed a p-value of 0.0. Therefore, reject the null hypothesis; there is a significant difference in the mean probability of successful summits occurring on commercial routes. Summit attempts on commercial routes are more likely to succeed.

3.3 SUMMARY

Data exploration revealed that most peaks are situated in China and Nepal, although all peaks are accessible through Nepal. The Khumbu-Rolwaling-Makalu region contains the most peaks. Most climbers are from the USA, Japan and Europe and the median age of climbers is 36 years. The age and gender distribution over time showed how, over the years, an increasing number of people, both men and women, are trekking the Himalayas. Lastly, not all members in an expedition successfully summits, even though the expedition is considered to have summitted successfully.

Supporting statistics confirmed what the correlation matrix suggested. Younger climbers, smaller expedition sizes, more hired guides and choosing commercial routes for summit attempts, all significantly contribute to the likelihood of summitting successfully. Interestingly, while attempting a summit today, you are more likely to succeed than at any time in the past.

Next, for a certain set of features that best describe yourself, the peak you are attempting and the summit approach, machine learning will predict your chances of success.

CHAPTER 4
DATA MODELLING

4.1 ANALYSIS

The data modelling approach included various models and methods to achieve a favourable accuracy for predicting summit success. The nature of the data lent itself to a supervised learning approach of classification to predict a single class output of success. Will a climber summit successfully or not?

4.2 METHOD

To predict summit success, first a simple heuristic estimator determined a baseline accuracy score, followed by more specialised estimators such as KNN, Logistic Regression, SVM, Random Forest and Gradient Boosting. The following sections describe each approach.

To measure the model effectiveness, the predictions from the test set were used to calculate model accuracy. The Receiver Operating Characteristic (ROC) score validated the performance of the classification model. Some models also produced a ROC curve for visual confirmation of model performance. Table 4.1 presents the scores for all model classifiers.

4.2.1 Heuristic

First the commercial route feature ‘comrte’ was directly used to predict summit success by comparing its binary values to that of the labelled data. This resulted in an accuracy of 0.6096.

Figure 4.1: ROC Curve for Heuristic Feature ‘mo2climb’

With the benefit of hindsight, the ‘feature importance’ parameter from the Random Forest classifier later revealed a better performing heuristic feature; the feature describing a member’s oxygen use while climbing, ‘mo2climb’. With the same approach as before, this feature was used as the heuristic feature with a baseline score of 0.7473 and a ROC of 0.6684 (ROC curve Figure 4.1).

4.2.2 KNN

As will be apparent later, all modelling approaches beat the heuristic model, however, the standard K-Nearest Neighbours (KNN) classifier “KNeighborsClassifier()” scored the worst of the advanced modelling approaches with an accuracy of 0.7625 and ROC score of 0.8071. Applying scaling and Principal Component Analysis (PCA) increased performance and accuracy by more than 3%. Figure 4.2 shows the ROC curve for the standard KNN model. This shows clear improvement over the heuristic approach.

Figure 4.2: ROC Curve for Standard KNN

4.2.3 Logistic Regression

Logistic Regression, “LogisticRegression()”, with scaling and a grid search for PCA and regularisation parameters, resulted in improvement over the standard KNN classifier, but did not meet the performance of the KNN classifier with scaling and PCA. Logistic Regression achieved an accuracy of 0.7743. See Table 4.1 for the best parameters that delivered this result.

4.2.4 SVM

The Support Vector Machine (SVM) model “SVC()” produced better results than both KNN and Logistic Regression. SVM with scaling and a grid search for PCA and regularisation parameters, resulted in accuracy of 0.8155.

4.2.5 Random Forest

The out-of-the-box application of the Random Forest classifiers “RandomForestClassifier()” and “ExtraTreesClassifier()” were the easiest to use, quickest to solve and nearly produced the best results. Both classifiers scored within 0.5% of each other. The slightly better of the two, RandomForestClassifier(), scored an accuracy of 0.8379 and a ROC score of 0.9089. Figure 4.3 shows further improvement in the ROC curve over the Standard KNN ROC curve (Figure 4.2).

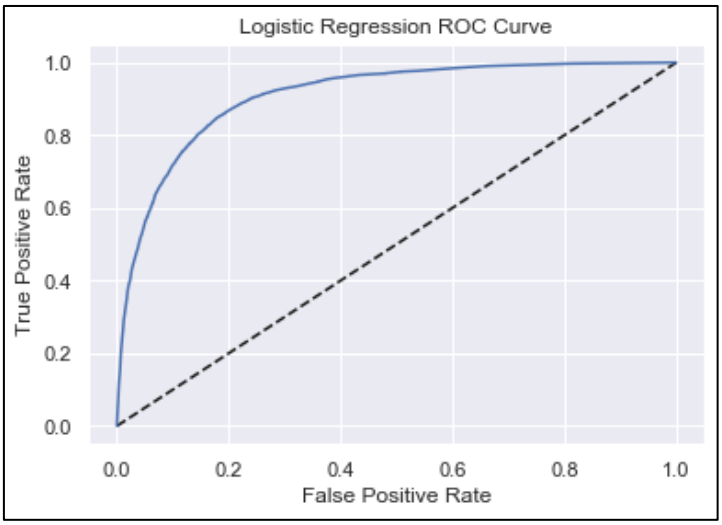

Figure 4.3: ROC Curve for Random Forest

4.2.6 Gradient Boosting

The Gradient Boosting model “GradientBoostingClassifier()” required more tuning and with grid search parameters, took the longest to solve, however, it produced the best results of all the classifiers tested. The standard model without grid search scored only 0.3% less. “AdaBoostClassifier()” scored worse and it’s performance is comparable to KNN with scaling and PCA. It scored 0.7895 for accuracy and 0.8599 for the ROC score.

The final score for Gradient Boosting with grid search, and best overall score, was 0.8454 for accuracy and 0.9188 for the ROC score. See Table 4.1 for the best parameters that delivered this result.

Figure 4.4 shows this final ROC curve. On careful inspection this curve covers a slightly larger area in the graph than the ROC curve for Random Forest (Figure 4.3). Unfortunately, the model learning parameters were not all stored separately after solving and therefore made a combined ROC plot impossible. Because some models take more than a day to solve, a combined ROC curve will only be included in future reports.

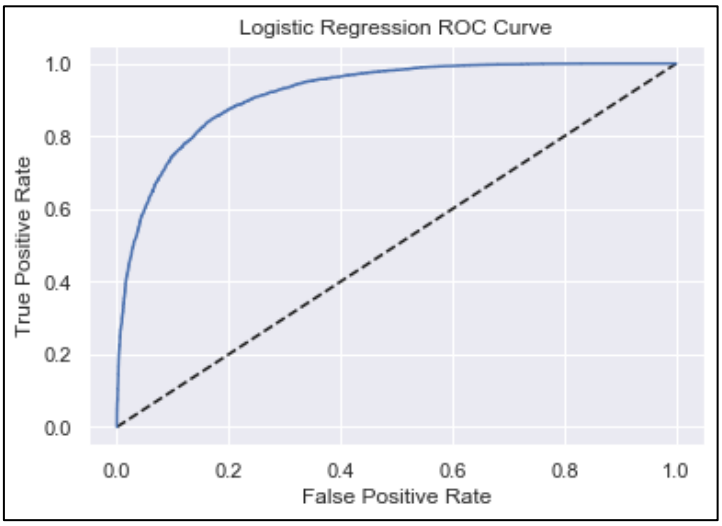

Figure 4.4: ROC Curve for Gradient Boost

4.3 RESULTS

Table 4.1: Model Results

Classifier					
No.	**Estimator**	**Detail**	**Accuracy**	**ROC**	**Best parameters**
1	**GradientBoostingClassifier with gridsearch**	**GridSearchCV: n_estimators=[10,50,100] learning_rate=[0.01,0.1,0.5,1] max_depth=[10,100]**	**0.8454**	**0.9188**	**n_estimators=100 learning_rate=0.1 max_depth=10**
2	GradientBoostingClassifier	n_estimators=50 learning_rate=0.5 max_depth=10	0.8427	0.9139	
3	RandomForestClassifier	n_estimators=100	0.8379	0.9089	
4	ExtraTreesClassifier	n_estimators=100	0.8310	0.8804	
5	SVM with gridsearch for scaling and PCA	C=[1,10,100] gamma=[0.1,0.01] n_components=0.95	0.8178		C=100 gamma=0.01 n_components=.95
6	KNeighborsClassifier with scaling	StandardScaler n_neighbours=3	0.7938	0.8439	
7	AdaBoostClassifier	n_estimators=100	0.7895	0.8599	
8	KNeighborsClassifier with scaling and PCA	StandardScaler PCA (n_components=0.8) n_neighbours=3	0.7887	0.8371	
9	LogisticRegression with gridsearch for scaling and PCA	GridSearchCV: C=log[-5:8:5] penalty=['l1','l2'] n_components=[1,.95,.9,.85]	0.7743		C=100e6 penalty='l2' n_components=.95
10	KNeighborsClassifier	n_neighbours=3	0.7625	0.8071	
11	Heuristic	mo2climb' feature	0.7473	0.6684	

The final tally ranking the best classifiers is tabulated in Table 4.1. The types of classifiers are grouped according to colour. Gradient Boost came out on top and is indicated by yellow, followed be Random Forest in green. KNN appears in the lower half of the table in amber. SVM, Logistic Regression and the Heuristic model, for no reason, appear as blue.

4.4 FINDINGS

Visualising the most important features (Figure 4.5) obtained from the Random Forest classifier revealed the most interesting results. Here the figure displays the features, in order of influence it has on the model. It is apparent that the age of a member (feature 'calcage') had the most significant influence in predicting the outcome of a summit attempt. This was hinted from the statistical inference completed before, but not how much influence compared to other features.

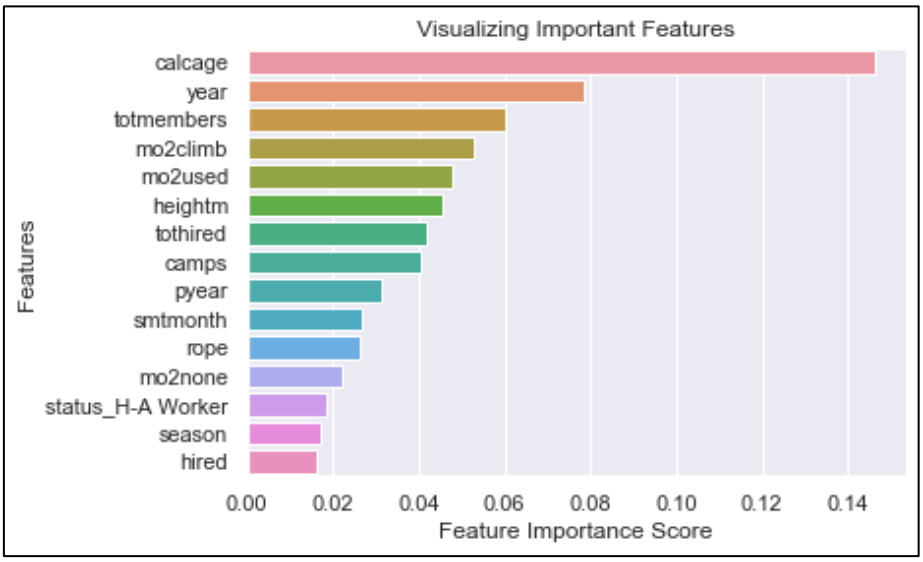

Figure 4.5: Important Features Ranking

Other features, in order of influence, are the year of summit attempt, expedition size, oxygen used by a member during the climb, any oxygen used by a member, the height of the peak and how many hired guides. The year of summit is only interesting for interest sake and has no bearing on predicting future summit success. There are of course more features to scrutinise, but it's influence on the outcome of summit success becomes nearly negligible.

In summary, for a certain individual with a plan to summit a peak in the Himalayas, Gradient Boosting with already determined optimum parameters, can with 85% certainty predict the outcome of the summit attempt.

CHAPTER 5
CONCLUSIONS

5.1 SUMMARY

Echoing insights gained from statistical inference with machine learning, the factors affecting a climber's chances of successful summiting a certain peak, has become progressively less uncertain. All things being equal, to significantly increase your chances of success it is important to consider the following statistics: Attempt the summit while you are still young. Join a smallest as possible expedition (with more hired guides). Take oxygen. Use it, especially in the climb. Possibly attempt lower summit peaks (not yet proven). Finally, hire as many guides as possible.

This advice is by no means exhaustive. There are more features that require analysis, such as, in order of importance; the height of the summit ('heightm'), number of camps, the year the summit route opened ('pyear') and the month when the summit occurred.

Regarding predictions, at the outset it was suggested that the model be used to predict summit success. For a certain set of features that best describe yourself, the peak you are attempting and the summit approach, machine learning would predict your success. Take the author as an example. A 38-year-old from South Africa attempting to summit Mount Everest (9000m), in year 2020, in the month of May, with 1600m fixed rope, plenty of oxygen for the climb and with five hired guides in a small group of six. Inputting this data in the Gradient Boosting classifier predicts a successful summit. However, simply leaving out oxygen use in the climb changes the outcome to failure.

5.2 FURTHER RESEARCH

Although the five most important features were analysed, there are still more important features in the top 10 worth analysing, such as summit height. Also, other features much lower down the importance hierarchy could be interesting in how they compare with each other. For example, within the nation (expedition country of origin) feature group, which country has more significance?

Considering unsuccessful summit attempts, multi-class labelled data exists that contains categorised reasons for the failed summit attempt. This could contribute further insight into why summit attempts fail.

Lastly, unsupervised learning could perhaps categorise occupational data into meaningful categories for inclusion in this model. This could reveal interesting occupation – success trends.

APPENDIX A:
ADDITIONAL DATA

Table A.1: Cleaned Data

Table:	**df**	Shape: 63113 x 51
Feature	**Description**	**Data type and status**
msuccess	Success (Yes/No)	Binary - Labelled data
season	Season	Float (cyclical - from categorical)
month (smtdate)	Date reached summit	Float (cyclical - from categorical)
nation	Principle nationality	Categorical (top 50)
status	Status	Categorical (top 50)
himal	Himal area	Categorical
region	Region	Categorical
phost	Peak host countries	Categorical
pstatus	Peak climbing status	Categorical
host	Host country	Categorical
heightm	Height (m)	Int
pyear	First ascent year	Int
year	Year	Int
camps	Number of high camps above BC	Int
rope	Amount of fixed rope (meters)	Int
totmembers	Number of members	Int
tothired	Number of hired personnel (above BC)	Int
calcage	Calculated age	Int
open	Peak open (Yes/No)	Binary
unlisted	Peak unlisted (Yes/No)	Binary
traverse	Traverse (Yes/No)	Binary
ski	Ski / snowboard descent (Yes/No)	Binary
parapente	Parapente descent (Yes/No)	Binary
nohired	No hired personnel used (above BC) (Yes/No)	Binary
o2used	Oxygen used (Yes/No)	Binary
o2none	Oxygen not used (Yes/No)	Binary
o2climb	Oxygen climbing (Yes/No)	Binary
o2descent	Oxygen descending (Yes/No)	Binary
o2sleep	Oxygen sleeping Yes/No)	Binary
o2medical	Oxygen used medically (Yes/No)	Binary
comrte	Commercial route (Yes/No)	Binary
stdrte	8000m standard route (Yes/No)	Binary
sex	Sex	Binary
leader	Leader (Yes/No)	Binary
deputy	Deputy leader (Yes/No)	Binary
bconly	BC / Advanced BC only (Yes/No)	Binary
disabled	Disabled (Yes/No)	Binary
hired	Hired local staff (Yes/No)	Binary
sherpa	Sherpa (Yes/No)	Binary

tibetan	Tibetan (Yes/No)	Binary
msolo	Solo (Yes/No)	Binary
mtraverse	Traverse (Yes/No)	Binary
mski	Ski / snowboard descent (Yes/No)	Binary
mparapente	Parapente descent (Yes/No)	Binary
mspeed	Speed ascent (Yes/No)	Binary
mo2used	Oxygen used (Yes/No)	Binary
mo2none	Oxygen not used (Yes/No)	Binary
mo2climb	Oxygen climbing (Yes/No)	Binary
mo2descent	Oxygen descending (Yes/No)	Binary
mo2sleep	Oxygen sleeping (Yes/No)	Binary
mo2medical	Oxygen used medically (Yes/No)	Binary

Source: Author, 2019.

Table A.2: 'Peaks' Table

Table:	**Peaks** (peaks.csv)	Shape: 468 x 22
Feature	**Description**	**Data type and status**
peakid	Peak ID	Dropped after merge
pkname	Peak name	Dropped
pkname2	Peak name 2	Dropped
location	Location	Dropped
heightm	Height (m)	Keep
heightf	Height (ft)	Dropped
himal	Himal area	Keep – categorical
region	Region	Keep – categorical
open	Peak open (Yes/No)	Keep
unlisted	Peak unlisted (Yes/No)	Keep
trekking	Trekking peak (Yes/No)	Dropped
trekyear	Trekking peak year	Dropped
restrict	Peak restrictions	Dropped
phost	Peak host countries	Keep – categorical
pstatus	Peak climbing status	Keep – categorical
pyear	First ascent year	Keep
pseason	First ascent season	Dropped
pexpid	First ascent expedition ID	Dropped
psmtdate	First ascent date	Dropped
pcountry	First ascent country	Dropped
psummiters	First ascent summiters	Dropped
psmtnote	First ascent comments	Dropped

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).

Table A.3: 'Peaks' – Categorical Features Description

Himal area	**Region**
0 – Unclassified	0 – Unclassified
1 – Annapurna	1 – Kangchenjunga-Janak
2 – Api/Byas Risi/Guras	2 – Khumbu-Rolwaling-Makalu
3 – Damodar	3 – Langtang-Jugal
4 – Dhaulagiri	4 – Manaslu-Ganesh
5 – Ganesh/Shringi	5 – Annapurna-Damodar-Peri
6 – Janak/Ohmi Kangri	6 – Dhaulagiri-Mukut
7 – Jongsang	7 – Kanjiroba-Far West
8 – Jugal	
9 – Kangchenjunga/Simhalila	**Peak host countries**
10 – Kanjiroba	0 – Unclassified
11 – Kanti/Palchung	1 – Nepal only
12 – Khumbu	2 – China only
13 – Langtang	3 – India only
14 – Makalu	4 – Nepal & China
15 – Manaslu/Mansiri	5 – Nepal & India
16 – Mukut/Mustang	6 – Nepal, China & India
17 – Nalakankar/Chandi/Changla	
18 – Peri	**Peak climbing status**
19 – Rolwaling	0 – Unknown
20 – Saipal	1 – Unclimbed
	2 – Climbed

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).

Table A.4: 'Expeditions' Features

Table:	**Expedition** (expeditions.csv)	Shape: 9959 x 65
Feature	**Description**	**Data type and status**
expid	Expedition ID	Dropped after merge
peakid	Peak ID	Dropped after merge
year	Year	Keep
season	Season	Keep – categorical/cyclical
host	Host country	Keep – categorical
route1	Climbing route 1	Dropped
route2	Climbing route 2	Dropped
route3	Climbing route 3	Dropped
route4	Climbing route 4	Dropped
nation	Principle nationality	Keep
leaders	Leadership	Dropped
sponsor	Expedition sponsor / name	Dropped
success1	Success on route 1 (Yes/No)	Dropped
success2	Success on route 2 (Yes/No)	Dropped
success3	Success on route 3 (Yes/No)	Dropped
success4	Success on route 4 (Yes/No)	Dropped
ascent1	Ascent numbers for route 1	Dropped
ascent2	Ascent numbers for route 2	Dropped
ascent3	Ascent numbers for route 3	Dropped
ascent4	Ascent numbers for route 4	Dropped
claimed	Success claimed (Yes/No)	Dropped
disputed	Success disputed (Yes/No)	Dropped
countries	Other countries	Dropped
approach	Approach march	Dropped
bcdate	Date arrived at base camp	Dropped
smtdate	Date reached summit	Keep, converted to month
smttime	Time reached summit	Dropped
smtdays	Nbr of days to summit / high-point	Dropped
totdays	Total number of days	Dropped
termdate	Date terminated	Dropped
termreason	Reason terminated	Dropped - multi class labelled
termnote	Termination details	Dropped
highpoint	Expedition high-point (m)	Dropped
traverse	Traverse (Yes/No)	Keep
ski	Ski / snowboard descent (Yes/No)	Keep
parapente	Parapente descent (Yes/No)	Keep
camps	Number of high camps above BC	Keep
rope	Amount of fixed rope (meters)	Keep
totmembers	Number of members	Keep
smtmembers	Number of members on summit	Dropped
mdeaths	Number of member deaths	Dropped
tothired	Number of hired personnel (above BC)	Keep
smthired	Number of hired personnel on summit	Dropped

hdeaths	Number of hired personnel deaths	Dropped
nohired	No hired personnel (above BC) (Yes/No)	Keep
o2used	Oxygen used (Yes/No)	Keep
o2none	Oxygen not used (Yes/No)	Keep
o2climb	Oxygen climbing (Yes/No)	Keep
o2descent	Oxygen descending (Yes/No)	Keep
o2sleep	Oxygen sleeping Yes/No)	Keep
o2medical	Oxygen used medically (Yes/No)	Keep
o2taken	Oxygen taken, not used (Yes/No)	Dropped
o2unkwn	Oxygen use unknown Yes/No)	Dropped
othersmts	Other summits	Dropped
campsites	Campsite details	Dropped
accidents	Accidents	Dropped
achievment	Achievements	Dropped
agency	Trekking agency	Dropped
comrte	Commercial route (Yes/No)	Keep
stdrte	8000m standard route (Yes/No)	Keep
primrte	Route info with primary exp (Yes/No)	Dropped
primmem	Mbr info with primary exp (Yes/No)	Dropped
primref	Literature info with primary exp (Yes/No)	Dropped
primid	Primary expedition ID (if any)	Dropped
chksum	Internal consistency check	Dropped

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).

Table A.5: 'Expeditions' – Categorical Features Description

Reason terminated	**Season**
0 – Unknown	0 – Unknown
1 – Success (main peak)	1 – Spring
2 – Success (subpeak)	2 – Summer
3 – Success (claimed)	3 – Autumn
4 – Bad weather (storms, high winds)	4 – Winter
5 – Bad conditions (deep snow, avalanching, falling ice, or rock)	
6 – Accident (death or serious injury)	**Host country**
7 – Illness, AMS, exhaustion, or frostbite	0 – Unknown
8 – Lack (or loss) of supplies or equipment	1 – Nepal
9 – Lack of time	2 – China
10 – Route technically too difficult, lack of experience, strength, or motivation	3 – India
11 – Did not reach base camp	
12 – Did not attempt climb	
13 – Attempt rumored	
14 – Other	

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).

Table A.6: 'Members' Features

Table:	**Members** (members.csv)	Shape: 65534 x 85
Feature	**Description**	**Data type and status**
expid	Expedition	Dropped after merge
membid	Expedition member	Dropped after merge
peakid	Peak ID	Dropped
myear	Year	Dropped
mseason	Season	Dropped
fname	First (given) name	Dropped
lname	Last (family) name	Dropped
sex	Sex	Keep
age	Age	Dropped
birthdate	Birth date	Dropped
yob	Year of birth	Dropped
calcage	Calculated age	Keep
citizen	Citizenship	Dropped
status	Status	Keep
residence	Residence (city / country)	Dropped
occupation	Occupation	Dropped
leader	Leader (Yes/No)	Keep
deputy	Deputy leader (Yes/No)	Keep
bconly	BC / Advanced BC only (Yes/No)	Keep
nottobc	Not to base camp (Yes/No)	Dropped
support	High-altitude support member (Yes/No)	Dropped
disabled	Disabled (Yes/No)	Keep
hired	Hired local staff (Yes/No)	Keep
sherpa	Sherpa (Yes/No)	Keep
tibetan	Tibetan (Yes/No)	Keep
msuccess	Success (Yes/No)	Keep - Labelled data
mclaimed	Success claimed (Yes/No)	Dropped
mdisputed	Success disputed (Yes/No)	Dropped
msolo	Solo (Yes/No)	Keep
mtraverse	Traverse (Yes/No)	Keep
mski	Ski / snowboard descent (Yes/No)	Keep
mparapente	Parapente descent (Yes/No)	Keep
mspeed	Speed ascent (Yes/No)	Keep
mhighpt	Expedition high-point reached (Yes/No)	Dropped
mperhighpt	Personal high-point	Dropped
msmtdate1	1st summit / high-point date	Dropped
msmtdate2	2nd summit date	Dropped
msmtdate3	3rd summit date	Dropped
msmttime1	1st summit / high-point time	Dropped
msmttime2	2nd summit time	Dropped
msmttime3	3rd summit time	Dropped
mroute1	1st ascent route	Dropped
mroute2	2nd ascent route	Dropped

mroute3	3rd ascent route	Dropped
mascent1	1st ascent number	Dropped
mascent2	2nd ascent number	Dropped
mascent3	3rd ascent number	Dropped
mo2used	Oxygen used (Yes/No)	Keep
mo2none	Oxygen not used (Yes/No)	Keep
mo2climb	Oxygen climbing (Yes/No)	Keep
mo2descent	Oxygen descending (Yes/No)	Keep
mo2sleep	Oxygen sleeping (Yes/No)	Keep
mo2medical	Oxygen used medically (Yes/No)	Keep
mo2note	Oxygen use reason	Dropped
death	Death (Yes/No)	Dropped
deathdate	Date of death	Dropped
deathtime	Time of death	Dropped
deathtype	Death type (cause)	Dropped
deathhgtm	Death height (m)	Dropped
deathclass	Death classification	Dropped
msmtbid	Summit Bid	Dropped
msmtterm	Summit bid termination reason	Dropped
hcn	Himalayan Club number	Dropped
mchksum	Internal consistency check	Dropped
host	Contained in 'expeditions.csv'	Dropped
comrte		Dropped
stdrte		Dropped
route1		Dropped
route2		Dropped
route3		Dropped
route4		Dropped
nation		Dropped
leaders		Dropped
sponsor		Dropped
termreason		Dropped
totmembers		Dropped
smtmembers		Dropped
mdeaths		Dropped
tothired		Dropped
nohired		Dropped
smthired		Dropped
hdeaths		Dropped
bcdate		Dropped
pkname	Contained in 'peaks.csv'	Dropped
heightm		Dropped

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).

Table A.7: 'Members' – Categorical Features Description

Summit bid termination reason
0 – Unspecified
1 – Success
2 – Success (subpeak)
3 – Bad weather (storms, high winds)
4 – Bad conditions (deep snow, avalanches, falling rock/ice)
5 – Accident (death or injury to self or others)
6 – Altitude (AMS symptoms, breathing or unwell)
7 – Exhaustion, fatigue, weakness or loss of motivation
8 – Frostbite, snowblindness or coldness
9 – Other illnesses or pains
10 – Lack of supplies or equipment problems
11 – O2 system failure
12 – Route difficulty, intimidation or insufficient ability
13 – Too late in day or too slow
14 – Assisting, guiding or accompanying others
15 – Route/camp preparation or rope fixing
16 – Insufficient time left for expedition
17 – Did not climb or intent to summit
18 – Other
19 – Unknown

Summit Bid
0 – Unspecified
1 – No summit bid
2 – Aborted below high camp
3 – Aborted at high camp
4 – Aborted above high camp
5 – Successful summit bid

Death type (cause)
0 – Unspecified
1 – AMS (acute mtn sickness)
2 – Exhaustion
3 – Exposure / frostbite
4 – Fall
5 – Crevasse
6 – Icefall collapse
7 – Avalanche
8 – Falling rock / ice
9 – Disappearance (unexplained)
10 – Illness (non-AMS)
11 – Other
12 – Unknown

Death classification
0 – Unspecified
1 – Death enroute BC
2 – Death at BC / ABC
3 – Route preparation
4 – Ascending in summit bid
5 – Descending from summit bid
6 – Expedition evacuation
7 – Other / Unknown

Source: Himalayan_Database_Guide.pdf, 2017 (edited by Author).