Machine Learning Exploration of Factors Affecting Student Retention

Contents

Introduction

This report aims to provide a thorough analysis of key factors influencing student retention within a further education (FE) provider. It supplements the insights from the 'Machine Learning Exploration of Factors Affecting Student Retention' Jupyter Notebook. This business case was chosen as retention is one of the primary measures in determining the quality of FE providers (Department for Education, 2024)

Data Retrieval

The anonymised dataset used was acquired from an FE provider. The below SQL (figure 1) was executed to collate the dataset:

Figure 1, SQL code for retrieving and anonymising the student retention data

```
WITH
    CauseForConcernComments
    AS
    (
        SELECT [redacted]StudentID, COUNT(*) AS ConcernCommentsCount
        FROM [redacted]Comments AS CMT
        WHERE CMT.academicyearID IN (
            '20/21'
            , '21/22'
            , '22/23'
        )
            AND CMT.Title LIKE '%Cause for Concern%'
        GROUP BY CMT.[redacted]StudentID
    )

SELECT DENSE_RANK() OVER (ORDER BY [redacted]) AS AnonymisedEnrolmentID
    , COALESCE( -- Get most relevant course delivery method
        E.DeliveryMethodID -- Get DeliveryMethodID first if not NULL
        , CASE WHEN [redacted] IN ([redacted]) THEN '01' -- Class Contact if these sites
        ELSE '07' END -- Anything else is Remote Learning
    ) AS CourseDeliveryMethodID
    , DENSE_RANK() OVER (ORDER BY [redacted]) AS AnonymisedSiteID
    , SUM(A.OverallPossibleAttendance) AS OverallPossibleAttendance
    , SUM(A.OverallAttendance) AS OverallAttendance
    , CCC.ConcernCommentsCount AS ConcernCommentsCount
    , C.CompletionStatusID
    , C.[Description] AS CompletionStatusDesc
FROM [redacted]students AS SD
    JOIN [redacted]Enrolment AS E
    ON [redacted]
    JOIN [redacted]CompletionStatus AS C
    ON [redacted]
    JOIN [redacted]Offering AS O
    ON [redacted]
    JOIN [redacted]attendance[redacted] AS A
    ON [redacted]
    LEFT JOIN CauseForConcernComments AS CCC
    ON [redacted]
WHERE SD.AcademicYearID IN ( -- Get three academic years of data
    '20/21'
    , '21/22'
    , '22/23'
    )
    AND C.[Description] NOT IN ( -- Exclude transferred and continuing enrolments
        'Transferred'
        , 'Continuing'
        )
    AND E.FundingID = [redacted] -- 16-19 EFA Funded Only
    AND O.OfferingTypeID = [redacted]-- Main courses only
    AND SD.isTestStudent IS NULL -- Exclude test students
    AND [redacted] NOT IN ( -- Exclude irrelevant college areas
            [redacted]
        )
GROUP BY
    [redacted]
    , E.DeliveryMethodID
   [redacted]
    , C.CompletionStatusID
    , C.[Description]
    , CCC.ConcernCommentsCount
    ;
```

The following columns were selected based on their relevance to student retention:

- AnonymisedEnrolmentID: An anonymised identifier for each enrolment.
- CourseDeliveryMethodID: An identifier for the delivery method of teaching.

- AnonymisedSiteID: An anonymised identifier representing the campus attended.
- OverallPossibleAttendance: The total possible attendance marks.
- OverallAttendance: The total lessons attended.
- ConcernCommentsCount: The total comments logged, pertaining to areas of concern.
- CompletionStatusID: The identifier of the students' completion status.
- CompletionStatusDesc: The description for the students' completion status.

Preliminary filtering was also necessary:

- Included three academic years to ensure an adequate and recent dataset.
- Excluded transfers and continuers, as these do not relate to finished courses.
- Included only 16-19 EFA students, as they constitute the majority of the college's cohort.
- Main courses only.
- Excluded test/irrelevant learners and college areas.

This data was converted into a CSV file and imported into the Jupyter Notebook as DataFrame ‘stud_ret_df’. Column 'OverallAttendancePercent' was created by dividing 'OverallAttendance' by 'OverallPossibleAttendance'.

Data Cleaning

During data cleaning, numerous steps were undertaken to ensure the dataset was prepared for analysis and machine learning. The shape of 'stud_ret_df' was verified to ensure it matched the CSV file. Printing the DataFrame information (figure 2) revealed several issues:

- 'ConcernCommentsCount' had 2116 missing values.
- 'ConcernCommentsCount' was type float, this was expected to be integer.
- 'CompletionStatusID' was type object, this was expected to be integer.

Figure 2, printing ‘stud_ret_df’ DataFrame information

```
1  stud_ret_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12044 entries, 0 to 12043
Data columns (total 9 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   AnonymisedEnrolmentID      12044 non-null  int64
 1   CourseDeliveryMethodID     12044 non-null  int64
 2   AnonymisedSiteID           12044 non-null  int64
 3   OverallPossibleAttendance  12044 non-null  int64
 4   OverallAttendance          12044 non-null  int64
 5   ConcernCommentsCount       9928 non-null   float64
 6   CompletionStatusID         12044 non-null  object
 7   CompletionStatusDesc       12044 non-null  object
 8   OverallAttendancePercent   12044 non-null  float64
dtypes: float64(2), int64(5), object(2)
```

Missing values in 'ConcernCommentsCount' were resolved by filling these with 0, as these represent learners with no comments. This allowed the column to be converted to integer.

'CompletionStatusID' contained 'X' values, which explained why this attribute was assigned as type object. The completion statuses relating to non-completers had significantly fewer instances compared to 'Completed' (figure 3).

Figure 3, value counts of student completion statuses

```
1  stud_ret_df[["CompletionStatusID","CompletionStatusDesc"]].value_counts()
```

```
CompletionStatusID  CompletionStatusDesc
2                   Completed               9566
3                   Withdrawn               2462
X                   Cancelled                 14
6                   Temporarily withdrawn      2
```

To address this, these were combined into 'CompletionStatusID' = 1 with 'CompletionStatusDesc' = 'Uncompleted'. This enabled the column to be converted to integer.

The value counts for 'CourseDeliveryMethodID' were printed, highlighting that only two categories existed, with '1' representing the vast majority (figure 4). Consequently, this column was dropped from the dataset.

Figure 4, value counts of 'CourseDeliveryMethodID'

```
stud_ret_df["CourseDeliveryMethodID"].value_counts()

CourseDeliveryMethodID
1     12005
7        39
```

The number of values for each 'AnonymisedSiteID' was displayed (figure 5), showing few enrolments at site '3'. Thus, these were filtered out.

Figure 5, value counts of 'AnonymisedSiteID'

```
stud_ret_df["AnonymisedSiteID"].value_counts()

AnonymisedSiteID
2     7909
1     4096
3       39
```

The DataFrame was checked for duplicate rows, none existed. Lastly, to enhance interpretability, the columns were renamed (figure 6)

Figure 6, renaming columns to improve interpretability

```
cols_rename_dict = {
    "AnonymisedEnrolmentID": "Anonymised ID",
    "AnonymisedSiteID": "Anonymised Site ID",
    "OverallPossibleAttendance": "Possible Attendance Marks",
    "OverallAttendance": "Actual Attendance Marks",
    "ConcernCommentsCount": "Comments of Concern",
    "CompletionStatusID": "Completion Status ID",
    "CompletionStatusDesc": "Completion Status Description",
    "OverallAttendancePercent": "Attendance%"
}

stud_ret_df = stud_ret_df.rename(columns=cols_rename_dict)
```

Exploratory Data Analysis (EDA)

EDA was conducted to understand the characteristics and relationships of the dataset. The statistics of 'stud_ret_df' were printed (figure 7)

Figure 7, first run of 'stud_ret_df' statistics

```
stud_ret_df.describe()
```

	Anonymised ID flo...	Anonymised Site ID f	Possible Attendan...	Actual Attendanc...	Comments of Con...	Completion Statu...	Attendance% float...
count	12005	12005	12005	12005	12005	12005	12005
mean	6003.683465	1.65880883	225.6303207	179.1985839	7.595835069	1.794169096	74.782717
std	3466.854687	0.4741291818	102.1332767	94.7614943	9.737332494	0.4043243257	22.56188596
min	1	1	0	0	0	1	0
25%	3002	1	177	108	1	2	64.51612903
50%	6003	2	257	200	4	2	81.25
75%	9004	2	296	252	10	2	91.44981413
max	12044	2	485	473	90	2	133.3333333

However, this revealed instances of over 100% attendance or 0 possible attendance, which were considered erroneous and filtered out. The statistics were reprinted thereafter (figure 8).

Figure 8, second run of 'stud_ret_df' statistics

```
stud_ret_df.describe()
```

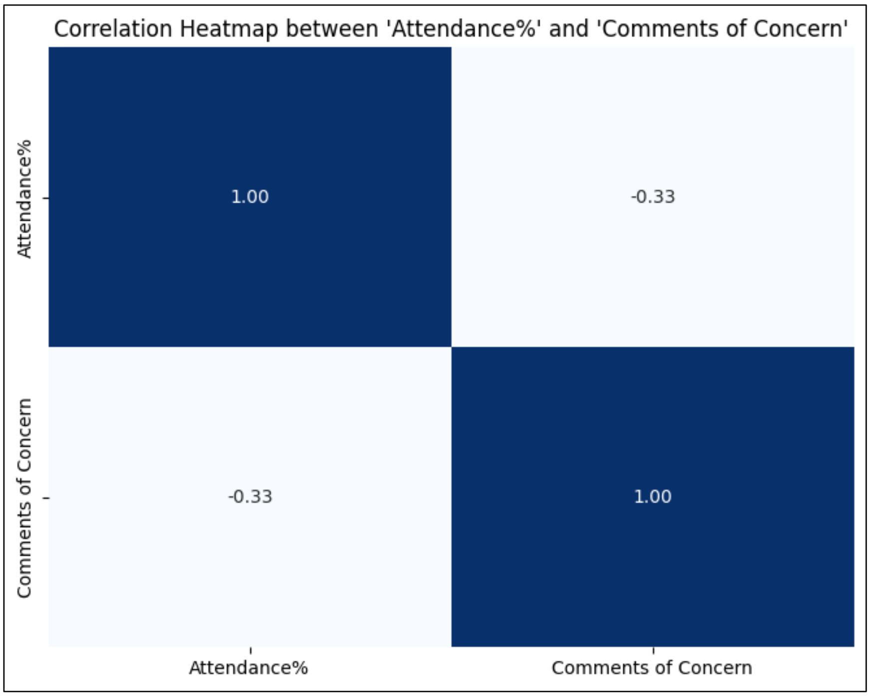
	Anonymised ID flo...	Anonymised Site ID f	Possible Attendan...	Actual Attendanc...	Comments of Con...	Completion Statu...	Attendance% float...
count	11759	11759	11759	11759	11759	11759	11759
mean	6022.328174	1.658899566	229.6416362	182.1901522	7.714856706	1.807636704	76.05707302
std	3466.651969	0.4740991906	98.43496625	92.40061519	9.793013965	0.3941736566	20.33183082
min	1	1	1	0	0	1	0
25%	3024.5	1	185	116	1	2	65.85855618
50%	6029	2	258	202	4	2	81.63934426
75%	9027.5	2	297	253	11	2	91.5432031
max	12044	2	485	473	90	2	100

This provided several observations:

- The average number of 'Comments of Concern' logged is 8.
- The quartiles and standard deviation for 'Comments of Concern' indicate that most students have fewer than 10 comments.
- The average 'Attendance%' is 76%

A heatmap was created to visualise the correlation between 'Attendance%' and 'Comments of Concern' (figure 9)

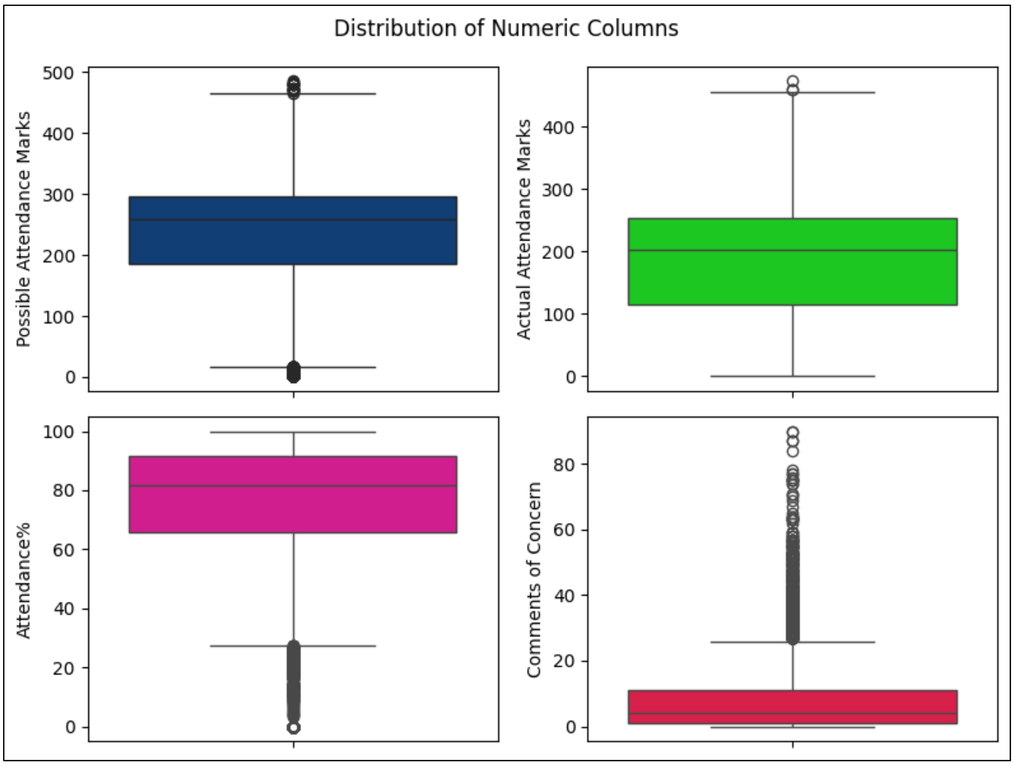
Figure 9, Correlation between 'Attendance%' and 'Comments of Concern'



This showed a moderate correlation between these attributes, which was considered promising for later machine learning analysis.

Following this, boxplots were generated to examine the distribution of the numeric columns (figure 10).

Figure 10, Numeric columns distribution



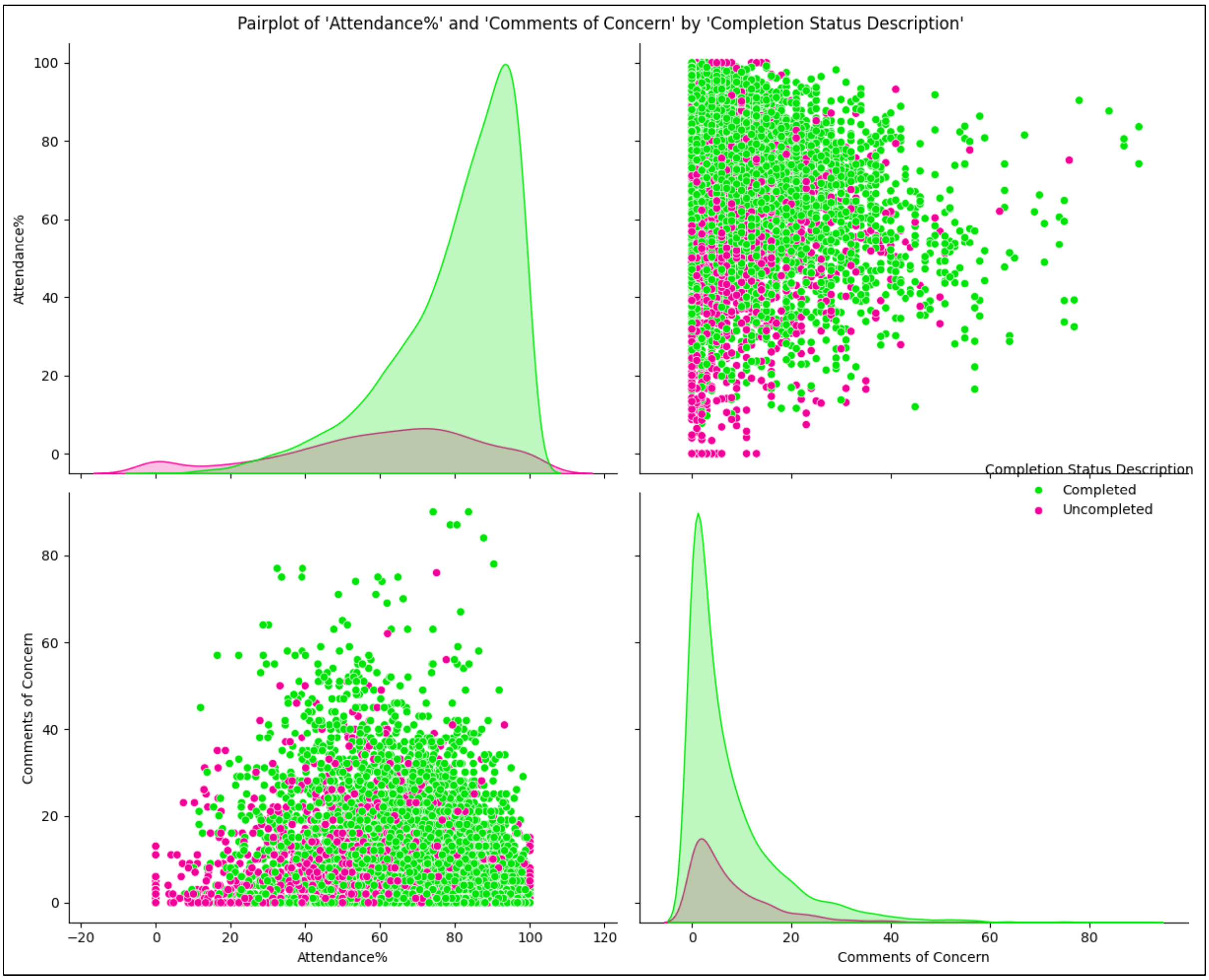
Numerous insights were revealed from this:

- In 'Possible Attendance Marks', minor outliers exist outside the lower and upper whiskers. This could be attributed to courses that are shorter/longer than usual.

- 'Attendance%' showed that most students have between 65% and 91% attendance.
- Outliers below the lower whisker in 'Attendance%' could indicate early withdrawers or rare attenders. These outliers were retained as they may prove valuable for predicting completion statuses.
- Significant outliers above the upper whisker in 'Comments of Concern' represent students with many concern comments. This is expected for recurring concerns, and thus, were retained.

Pair plots were created to compare 'Attendance%' with 'Comments of Concern', by 'Completion Status Description' (figure 11).

Figure 11, comparing 'Attendance%' with 'Comments of Concern', by 'Completion Status Description'

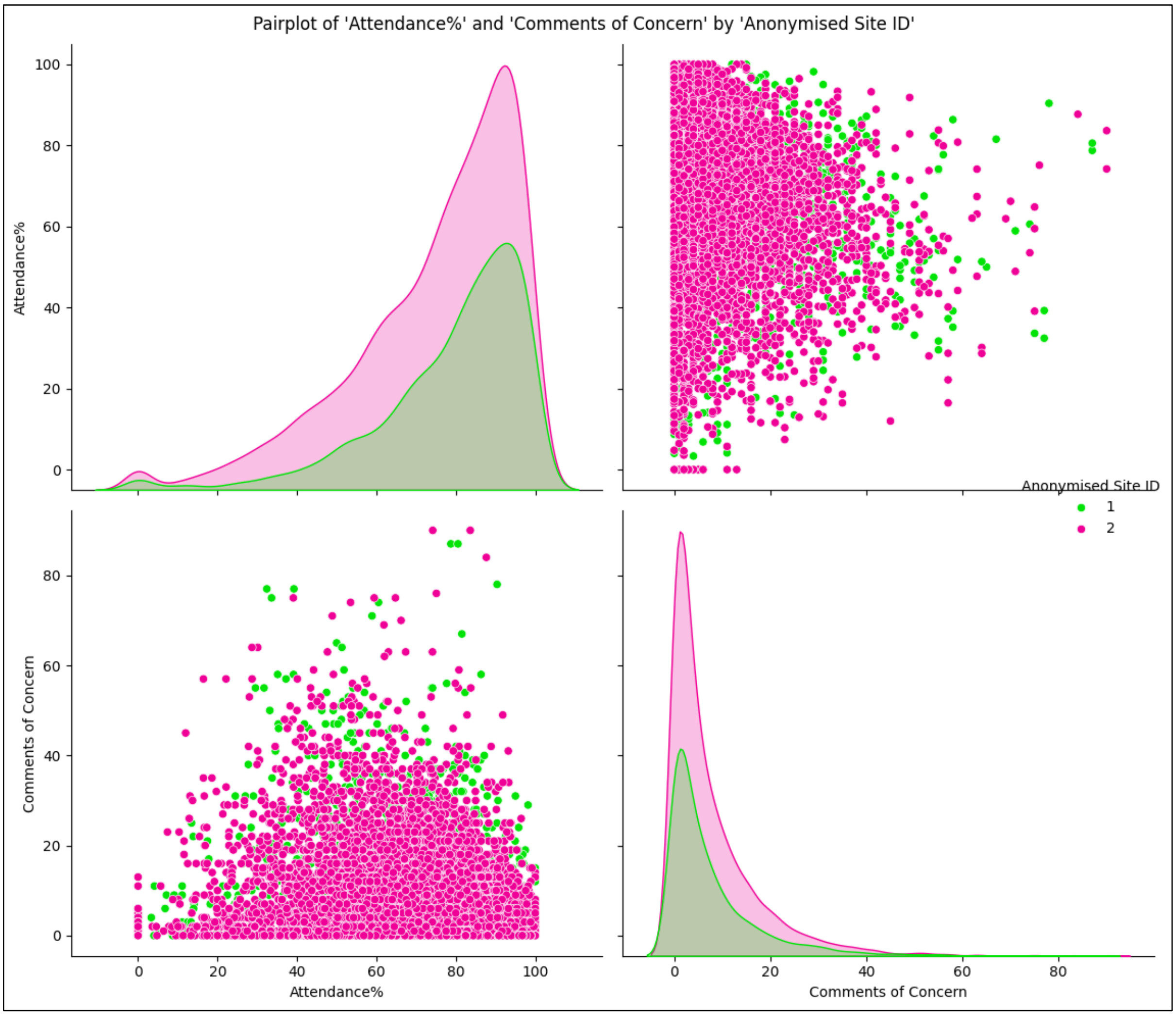


The following observations were made:

- The density of high attenders appears significantly higher for completers compared to non-completers (top-left).
- A slight trend is observed, with completers having higher attendance and fewer comments of concern (bottom-left).
- Completers exhibit a higher density of comments of concern; however, this is likely influenced by the fact that around 80% of the students are completers (bottom-right).

A similar pair plot was generated, this time grouping by 'Anonymised Site ID' (figure 12).

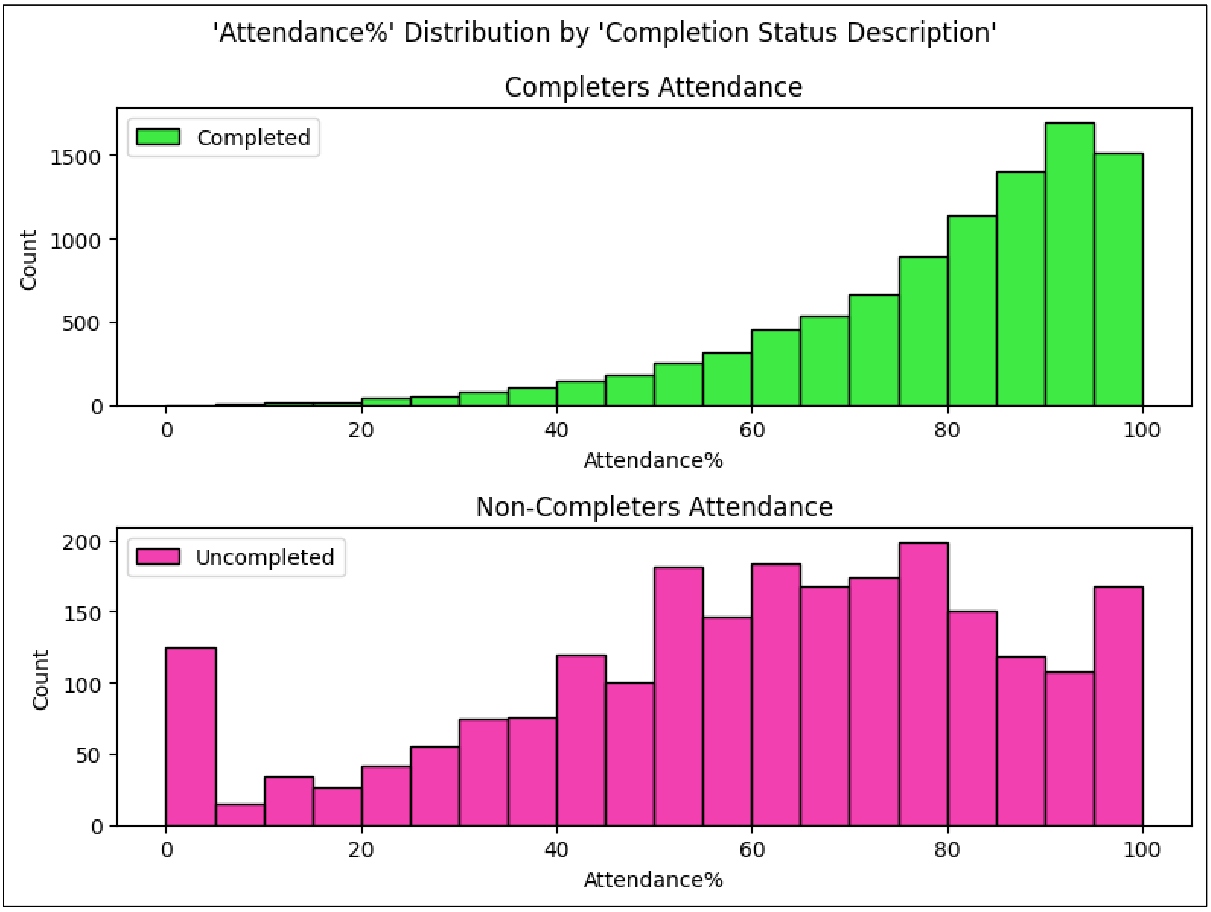
Figure 12, comparing 'Attendance%' with 'Comments of Concern', by ' Anonymised Site ID '



Site 2 shows a higher density of both higher attendance and comments of concern. However, this could be attributed to having a larger student population (top-left and bottom-right). There is also no noticeable difference in the concentration of comments of concern between the two sites (bottom-left).

Histograms were created to analyse the distribution of 'Attendance%' by 'Completion Status Description' (figure 13)

Figure 13, distribution of 'Attendance%' by 'Completion Status Description'

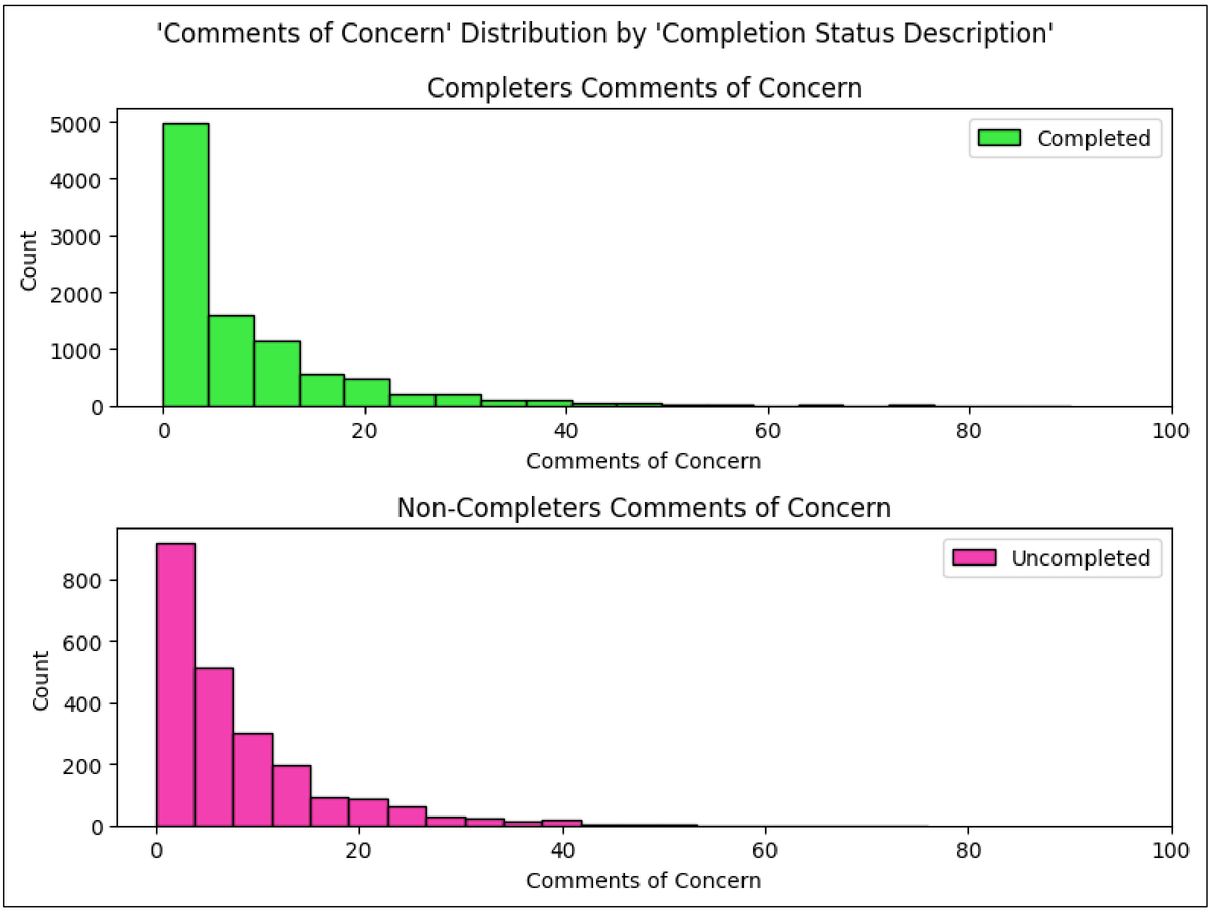


The following insights were noted:

- The distribution is vastly different, with completers exhibiting a left-skewed distribution and non-completers showing a loosely normal distribution.
- Most completers have attendance between 80-100%, whereas non-completers are concentrated between 50-80%.
- A notable concentration of non-completers is observed with close to 0% attendance.

A different histogram was generated to visualise the distribution of 'Comments of Concern' by 'Completion Status Description' (figure 14).

Figure 14, distribution of ‘Comments of Concern’ by ‘Completion Status Description’



This facilitated the following observations:

- The distribution for both statuses is very similar, both showing a right-skewed distribution.
- Non-completers tend to have a somewhat higher distribution of more concern comments compared to completers.

Machine Learning Modelling

Machine learning models were employed to make predictions and enhance the understanding of trends associated with student retention.

Predicting Students' Completion Statuses

Support Vector Classification (SVC) was utilised to predict students' completion statuses. The initial run intended to evaluate the behaviour of machine learning with the dataset and identify issues. This preliminary assessment was conducted using stratified k-fold cross-validation to ensure balanced splits of the data (Prusty, Patnaik, and Dash, 2022).

For the classification models, classification reports were generated, and confusion matrices were produced to summarise the performance (Agarwal, et al, 2021) (figure 15).

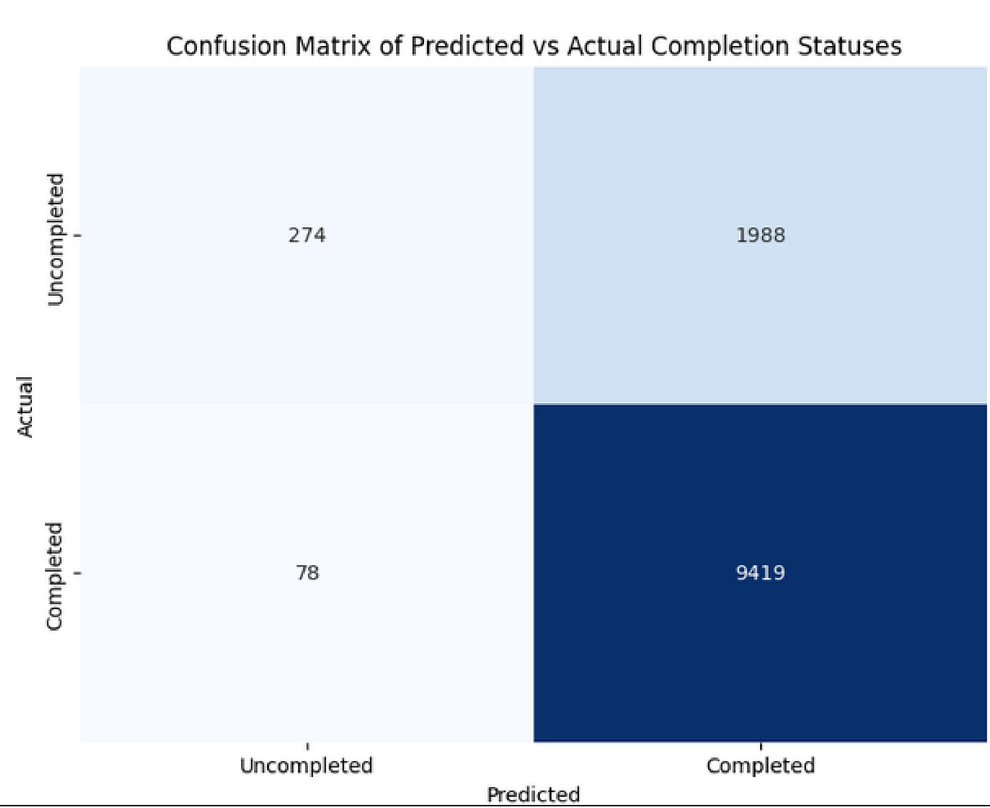
Figure 15, initial model evaluation results

```
Cross-validation results:  [0.82270408 0.82993197 0.82227891 0.8252551  0.82135262]
Mean cross-validation result:  0.8243045367870672

Classification report:
              precision    recall  f1-score   support

           1       0.78      0.12      0.21      2262
           2       0.83      0.99      0.90      9497

    accuracy                           0.82     11759
   macro avg       0.80      0.56      0.56     11759
weighted avg       0.82      0.82      0.77     11759
```



An accuracy score of 82% appeared promising. However, further inspection revealed low recall, f1-score, and incorrect predictions for uncompleted learners, indicating that the model was mainly effective in predicting completed learners. This correlates with the fact that approximately 80% of learners completed their course.

To identify the optimal SVC model, three kernels and regularisation (C) values were tested (table 1). To address the imbalanced completion statuses, SMOTE was implemented to oversample non-completers (Chawla, et al, 2002).

Table 1, testing SVC kernels and C-values

Kernel: linear, C-value: 0.1 Cross-validation results: [0.76036132 0.77417641 0.7698033 0.76289208 0.75598086] Mean cross-validation result: 0.7646427923736498 Classification report: precision recall f1-score support 1 0.42 0.55 0.48 473 2 0.88 0.81 0.84 1879 accuracy 0.76 2352 macro avg 0.65 0.68 0.66 2352 weighted avg 0.79 0.76 0.77 2352
Kernel: linear, C-value: 1 Cross-validation results: [0.76036132 0.77311371 0.7698033 0.76342371 0.7549176] Mean cross-validation result: 0.7643239261003119 Classification report: precision recall f1-score support 1 0.42 0.55 0.48 473 2 0.88 0.81 0.84 1879 accuracy 0.76 2352 macro avg 0.65 0.68 0.66 2352 weighted avg 0.79 0.76 0.77 2352
Kernel: linear, C-value: 10 Cross-validation results: [0.76036132 0.77311371 0.7698033 0.76342371 0.7549176] Mean cross-validation result: 0.7643239261003119 Classification report: precision recall f1-score support 1 0.42 0.55 0.48 473 2 0.88 0.81 0.84 1879 accuracy 0.76 2352 macro avg 0.65 0.68 0.66 2352 weighted avg 0.79 0.76 0.77 2352
Kernel: rbf, C-value: 0.1 Cross-validation results: [0.76567481 0.75876727 0.76767677 0.76289208 0.75704413] Mean cross-validation result: 0.762411010942808 Classification report:

```
          precision    recall  f1-score   support

       1       0.41      0.56      0.47       473
       2       0.88      0.80      0.84      1879

    accuracy                           0.75      2352
   macro avg       0.64      0.68      0.65      2352
weighted avg       0.78      0.75      0.76      2352
```

```
Kernel: rbf, C-value: 1
Cross-validation results:  [0.75292242 0.74176408 0.76076555 0.74960128
0.74428495]
Mean cross-validation result:  0.7498676569374034

Classification report:
          precision    recall  f1-score   support

       1       0.40      0.59      0.48       473
       2       0.88      0.78      0.83      1879

    accuracy                           0.74      2352
   macro avg       0.64      0.68      0.65      2352
weighted avg       0.79      0.74      0.76      2352
```

```
Kernel: rbf, C-value: 10
Cross-validation results:  [0.74601488 0.73113709 0.7533227  0.73418394
0.73365231]
Mean cross-validation result:  0.7396621847989374

Classification report:
          precision    recall  f1-score   support

       1       0.39      0.62      0.48       473
       2       0.89      0.76      0.82      1879

    accuracy                           0.73      2352
   macro avg       0.64      0.69      0.65      2352
weighted avg       0.79      0.73      0.75      2352
```

```
Kernel: poly, C-value: 0.1
Cross-validation results:  [0.79702444 0.80286929 0.80382775 0.79160021
0.79585327]
Mean cross-validation result:  0.798234992692177

Classification report:
          precision    recall  f1-score   support

       1       0.46      0.40      0.43       473
       2       0.85      0.89      0.87      1879

    accuracy                           0.79      2352
   macro avg       0.66      0.64      0.65      2352
weighted avg       0.77      0.79      0.78      2352
```

```
Kernel: poly, C-value: 1
Cross-validation results:  [0.79702444 0.80499469 0.80329612 0.79213184
0.79425837]
Mean cross-validation result:  0.7983410931282735

Classification report:
              precision    recall  f1-score   support

           1       0.47      0.40      0.43       473
           2       0.85      0.89      0.87      1879

    accuracy                           0.79      2352
   macro avg       0.66      0.64      0.65      2352
weighted avg       0.78      0.79      0.78      2352
```

```
Kernel: poly, C-value: 10
Cross-validation results:  [0.79702444 0.80499469 0.80276449 0.79266348
0.79479001]
Mean cross-validation result:  0.7984474195503894

Classification report:
              precision    recall  f1-score   support

           1       0.47      0.39      0.43       473
           2       0.85      0.89      0.87      1879

    accuracy                           0.79      2352
   macro avg       0.66      0.64      0.65      2352
weighted avg       0.78      0.79      0.78      2352
```

The polynomial kernel yielded the best results. Thus, the final SVC model, employing this kernel with a C-value of 1, was created (figure 16).

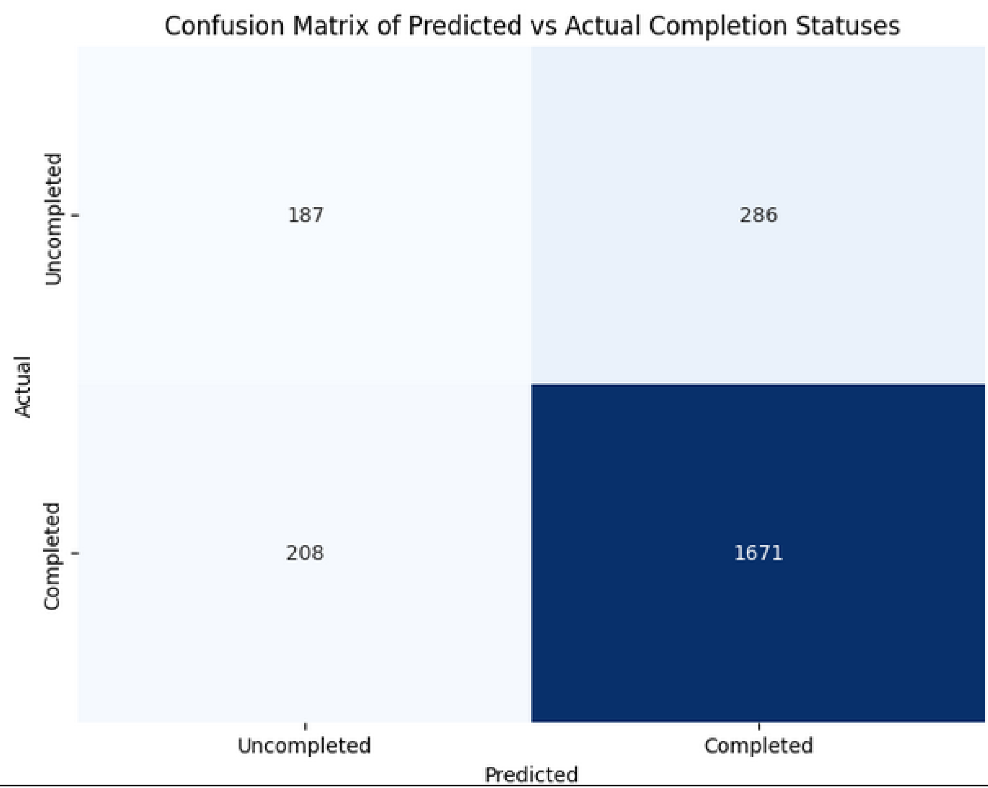
Figure 16, final SVC model evaluation results

```
Cross-validation results:  [0.79702444 0.80499469 0.80329612 0.79213184 0.79425837]
Mean cross-validation result:  0.7983410931282735

Classification report:
              precision    recall  f1-score   support

           1       0.47      0.40      0.43       473
           2       0.85      0.89      0.87      1879

    accuracy                           0.79      2352
   macro avg       0.66      0.64      0.65      2352
weighted avg       0.78      0.79      0.78      2352
```



Despite a slight reduction in accuracy compared to the initial run, this model demonstrated improved performance in predicting non-completers. This is evidenced by higher precision, recall, f1-score, and fewer incorrect predictions of non-completers.

Logistic regression was also utilised to predict completion statuses. Three solvers and regularisation values were tested (table 2).

Table 2, testing logistic regression solvers and C-values

Solver: newton-cg, C-value: 0.1 Cross-validation results: [0.75026567 0.76036132 0.74906964 0.74534822 0.73577884] Mean cross-validation result: 0.7481647392884039 Classification report: precision recall f1-score support 1 0.41 0.60 0.48 473 2 0.88 0.78 0.83 1879 accuracy 0.74 2352 macro avg 0.65 0.69 0.66 2352 weighted avg 0.79 0.74 0.76 2352
Solver: newton-cg, C-value: 1 Cross-validation results: [0.75026567 0.76036132 0.74800638 0.74641148 0.73684211] Mean cross-validation result: 0.7483773921326357 Classification report: precision recall f1-score support 1 0.41 0.60 0.48 473 2 0.88 0.78 0.83 1879 accuracy 0.74 2352 macro avg 0.65 0.69 0.66 2352 weighted avg 0.79 0.74 0.76 2352
Solver: newton-cg, C-value: 10 Cross-validation results: [0.75026567 0.76036132 0.74800638 0.74641148 0.73684211] Mean cross-validation result: 0.7483773921326357 Classification report: precision recall f1-score support 1 0.41 0.60 0.48 473 2 0.88 0.78 0.83 1879 accuracy 0.74 2352 macro avg 0.65 0.69 0.66 2352 weighted avg 0.79 0.74 0.76 2352
Solver: lbfgs, C-value: 0.1 Cross-validation results: [0.75026567 0.76036132 0.74906964 0.74534822 0.73577884] Mean cross-validation result: 0.7481647392884039 Classification report: precision recall f1-score support

1 0.41 0.60 0.48 473
2 0.88 0.78 0.83 1879

accuracy 0.74 2352
macro avg 0.65 0.69 0.66 2352
weighted avg 0.79 0.74 0.76 2352

Solver: lbfgs, C-value: 1
Cross-validation results: [0.75026567 0.76036132 0.74800638 0.74641148 0.73684211]
Mean cross-validation result: 0.7483773921326357

Classification report:
precision recall f1-score support

1 0.41 0.60 0.48 473
2 0.88 0.78 0.83 1879

accuracy 0.74 2352
macro avg 0.65 0.69 0.66 2352
weighted avg 0.79 0.74 0.76 2352

Solver: lbfgs, C-value: 10
Cross-validation results: [0.75026567 0.76036132 0.74800638 0.74641148 0.73684211]
Mean cross-validation result: 0.7483773921326357

Classification report:
precision recall f1-score support

1 0.41 0.60 0.48 473
2 0.88 0.78 0.83 1879

accuracy 0.74 2352
macro avg 0.65 0.69 0.66 2352
weighted avg 0.79 0.74 0.76 2352

Solver: liblinear, C-value: 0.1
Cross-validation results: [0.74973433 0.76036132 0.74906964 0.74587985 0.73631047]
Mean cross-validation result: 0.7482711222070246

Classification report:
precision recall f1-score support

1 0.41 0.60 0.48 473
2 0.88 0.78 0.83 1879

accuracy 0.74 2352
macro avg 0.65 0.69 0.66 2352
weighted avg 0.79 0.74 0.76 2352

Solver: liblinear, C-value: 1

```
Cross-validation results:  [0.75026567 0.76036132 0.74800638 0.74641148
0.73684211]
Mean cross-validation result:  0.7483773921326357

Classification report:
              precision    recall  f1-score   support

           1       0.41      0.60      0.48       473
           2       0.88      0.78      0.83      1879

    accuracy                           0.74      2352
   macro avg       0.65      0.69      0.66      2352
weighted avg       0.79      0.74      0.76      2352
```

```
Solver: liblinear, C-value: 10
Cross-validation results:  [0.75026567 0.76036132 0.74800638 0.74641148
0.73684211]
Mean cross-validation result:  0.7483773921326357

Classification report:
              precision    recall  f1-score   support

           1       0.41      0.60      0.48       473
           2       0.88      0.78      0.83      1879

    accuracy                           0.74      2352
   macro avg       0.65      0.69      0.66      2352
weighted avg       0.79      0.74      0.76      2352
```

Although minimal differences were observed, newton-cg solver with C-value 1 were chosen for the parameters of the final model (figure 17).

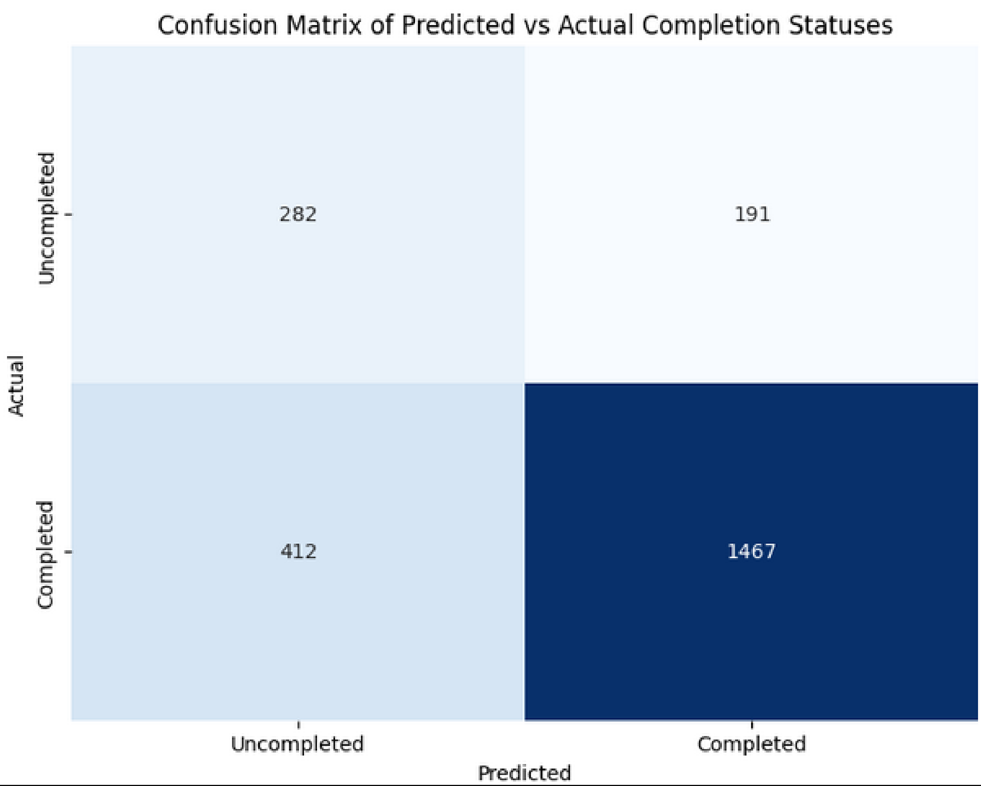
Figure 17, final logistic regression model evaluation results

```
Cross-validation results:  [0.75026567 0.76036132 0.74800638 0.74641148 0.73684211]
Mean cross-validation result:  0.7483773921326357

Classification report:
              precision    recall  f1-score   support

           1       0.41      0.60      0.48       473
           2       0.88      0.78      0.83      1879

    accuracy                           0.74      2352
   macro avg       0.65      0.69      0.66      2352
weighted avg       0.79      0.74      0.76      2352
```

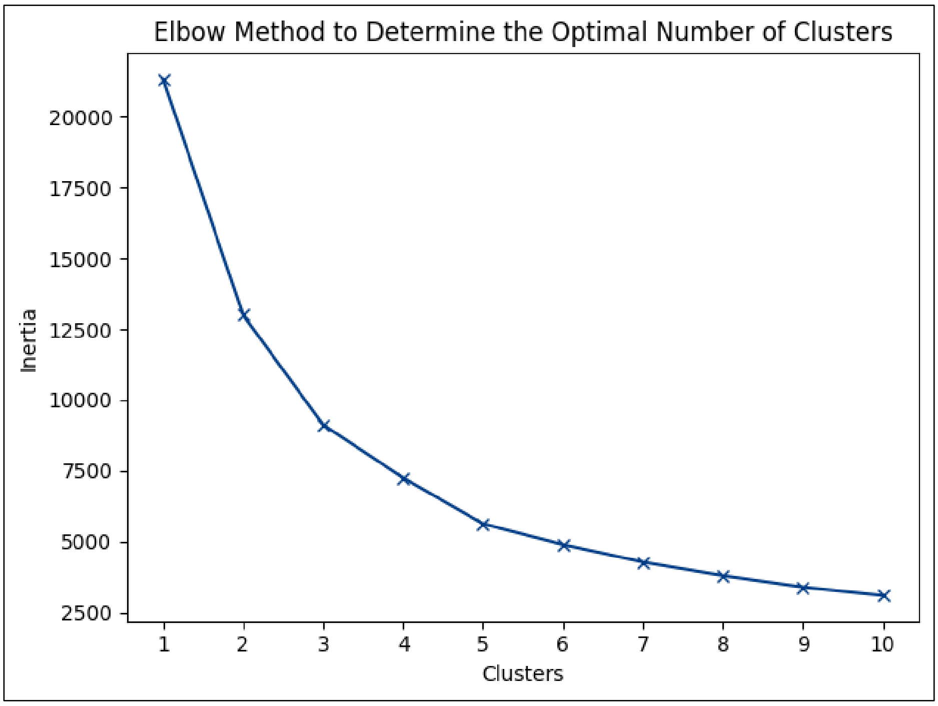


In terms of performance, this model identified non-completers slightly better than the SVC model. However, its overall accuracy was lower.

Identifying Student Groups via Clustering

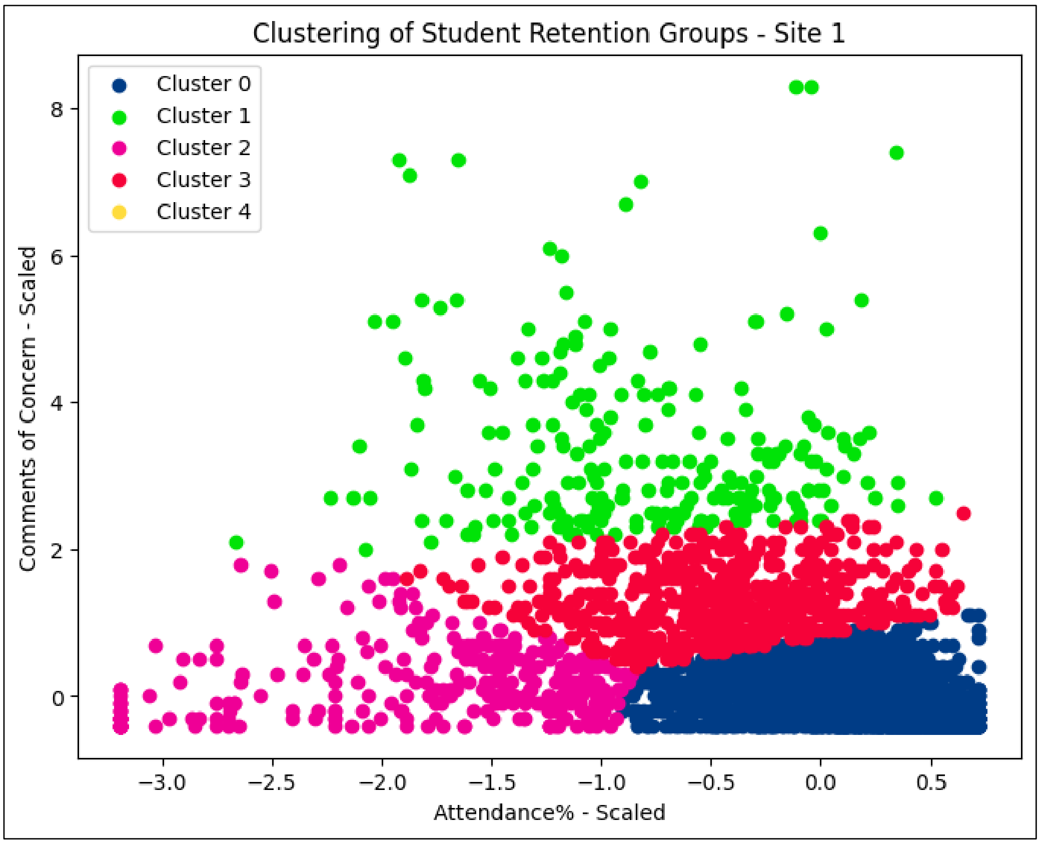
A k-means clustering model was implemented to identify student groups within the dataset. To determine the ideal number of clusters, the elbow method was employed (Nainggolan, et al, 2019) (figure 18).

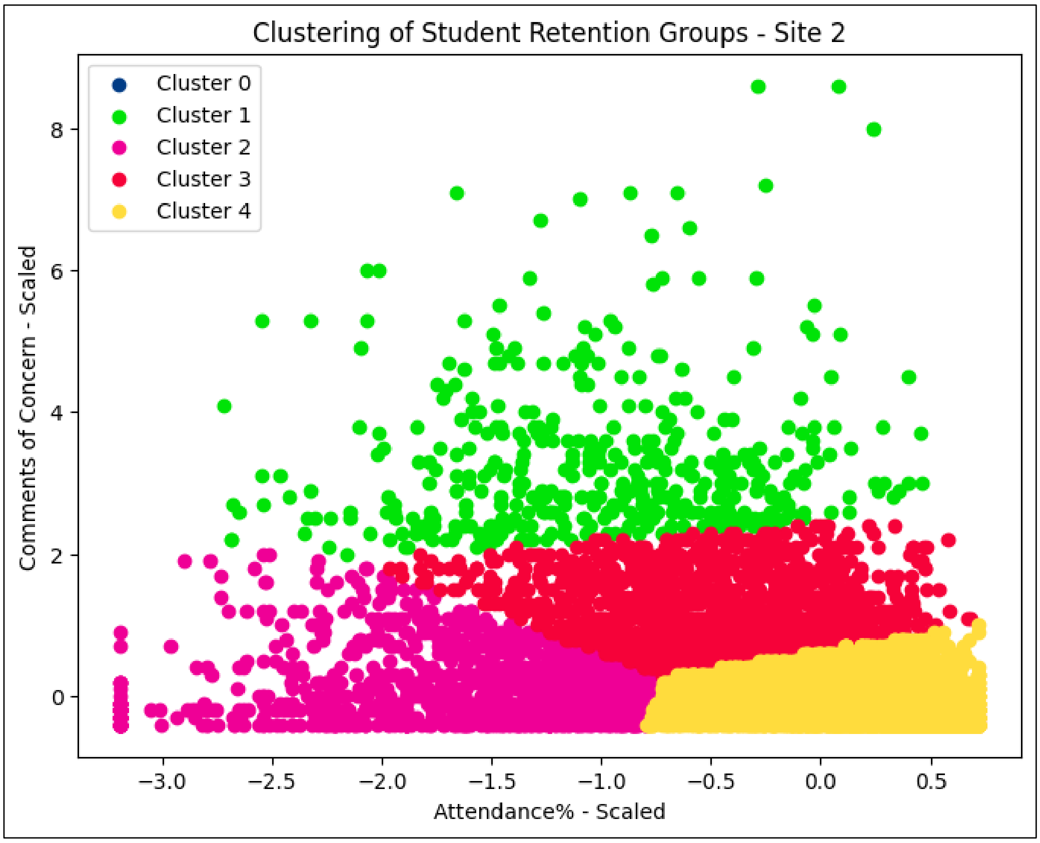
Figure 18, determining the optimal number of clusters using the elbow method



5 clusters were chosen from this. 2D scatter plots of the clusters were generated for each site (figure 19).

Figure 19, visualisation of clustered student retention groups for each site





This revealed five distinct student groups:

- Cluster 0 & 4: High attenders with few concern comments. These might be different clusters due to differing student population at each site.
- Cluster 1: Students with many concern comments and varying attendance.
- Cluster 2: Low attenders with minimal concern comments.
- Cluster 3: Average attenders and concern comments.

Estimating 'Comments of Concern'

A linear regression model was developed to estimate 'Comments of Concern' based on 'Attendance%'. The model's performance was evaluated with regression scoring metrics, using cross-validation with repeated k-fold across different data partitions (Wong and Yeh, 2019) (figure 20).

Figure 20, linear regression model evaluation results

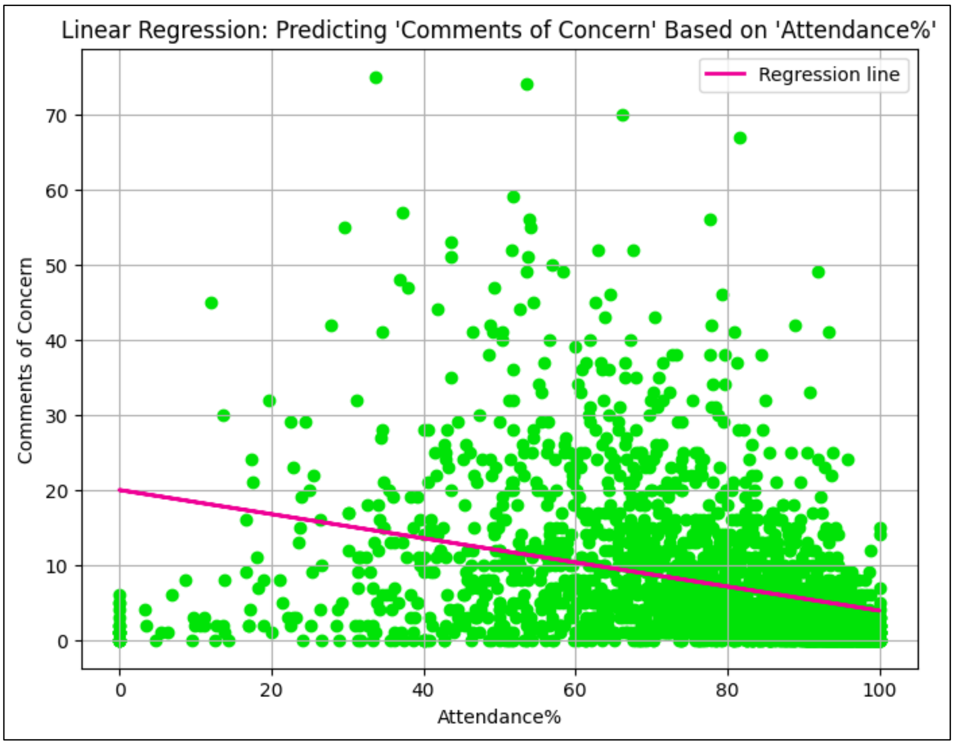
```
Intercept: 19.966698434456436
Coefficient: -0.16069660944994513

Cross-validation R2: [0.10997237 0.09690809 0.12088992 0.08544401 0.131708   0.09973364
 0.10986309 0.131638   0.11136975 0.09419197 0.11403141 0.10170729
 0.11811599 0.08291723 0.13250405 0.11147782 0.11540517 0.10828529
 0.0932905  0.12160604 0.09321851 0.12701677 0.13158275 0.1115895
 0.08405928]
Mean cross validation R2: 0.10954105766961483

Cross-validation mean squared error: [ -92.40438842  -84.1848763   -81.15901937  -86.94415396  -83.6738768
  -83.32571801  -77.70113985  -83.70202385  -92.01324643  -91.52485718
  -87.1911184   -82.27278122  -92.43140121  -91.48264777  -74.90575926
  -86.03228754  -87.77749633  -78.95149607  -89.68642837  -85.63014363
 -100.05028424  -83.95422006  -81.24320799  -75.64154477  -87.41758791]
Mean cross validation mean squared error: -85.65206819740389

Cross-validation root mean squared error: [ -9.61272014  -9.17523168  -9.00883008  -9.32438491  -9.14734261
  -9.12829217  -8.814825    -9.14888102  -9.59235354  -9.56686245
  -9.33761845  -9.07043446  -9.61412509  -9.56465618  -8.65481134
  -9.27535916  -9.36896453  -8.88546544  -9.47029188  -9.25365569
 -10.0025139   -9.16265355  -9.01350143  -8.69721477  -9.34973732]
Mean cross validation root mean squared error: -9.249629070707684

Cross-validation mean absolute error: [-6.40387947 -6.23718584 -6.27525882 -6.23690975 -6.14509112 -6.16989389
 -6.0630786  -6.20244678 -6.45042392 -6.40708467 -6.25059181 -6.2364318
 -6.42595245 -6.31676406 -6.06252509 -6.12976381 -6.34754406 -6.26825606
 -6.2174912  -6.32076622 -6.57007667 -6.22014579 -6.20357298 -6.10029978
 -6.19715744]
Mean cross validation mean absolute error: -6.258343683054722
```



The performance of this model was quite poor:

- Loose-Fitting Regression Line: The regression line had a loose fit with many points deviating significantly. The intercept of approximately 19.97 and the coefficient of -0.16 show a weak inverse relationship.
- Low R^2 Score: The model's mean R^2 score of 0.11 indicates that 'Attendance%' explains only a small fraction of the variance in 'Comments of Concern'.
- Error Metrics: The RMSE and MAE show that the model’s predictions were off by 6-9 concern comments on average.

Concluding Remarks

Overall, the analysis achieved moderate success in examining the effects of attendance, concern comments, and site on student retention. The final SVC model demonstrated the highest effectiveness in predicting learners' completion statuses, achieving an overall accuracy of 79.8%. Hence, this model can offer valuable insights for FE providers in estimating retention rates, however it may not be precise enough for definitive figures, as it cannot account for individual circumstances such as unexpected withdrawals for personal reasons.

The k-means clustering model identified five student groups, which could be beneficial in tailoring to student needs based on their cluster group. For instance, additional support could be provided to students in cluster 1, characterised by high concern comments.

In predicting the number of concern comments based on attendance, the analysis revealed that the correlation between these attributes was insufficient for reliable predictions, as shown by the poor performance of the linear regression model.

References

Agarwal, A., et al (2021) 'Classification model for accuracy and intrusion detection using machine learning approach', PeerJ Computer Science, 7 [online]. Available at: https://peerj.com/articles/cs-437/ (Accessed 8th June 2024)

Chawla, N.V., et al (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, 16 [online]. Available at: https://www.jair.org/index.php/jair/article/view/10302 (Accessed 8th June 2024)

Department for Education (2024) *Qualification achievement rates: Business Rules 2023 to 2024* [Online]. Available at: https://assets.publishing.service.gov.uk/media/66597c5116cf36f4d63ebc98/Qualification_achievement_rates_Business_Rules_2023_to_2024.pdf (Accessed 6th June 2024)

Nainggolan, R., et al (2019) 'Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method'. *Journal of Physics: Conference Series,* Vol. 1361, No. 1 [online]. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1361/1/012015/meta (Accessed 8th June 2024)

Prusty, S., Patnaik, S., and Dash, S.K. (2022) 'SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer', *Frontiers in Nanotechnology*, 4 [online]. Available at: https://www.frontiersin.org/articles/10.3389/fnano.2022.972421/full (Accessed 8th June 2024)

Wong, T.T. and Yeh, P.Y. (2019) 'Reliable accuracy estimates from k-fold cross validation', *IEEE Transactions on Knowledge and Data Engineering*, *32*(8) [online]. Available at: https://ieeexplore.ieee.org/abstract/document/8698831 (Accessed 8th June 2024)