Imports

In [1]:
```
import numpy as np
import math
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [2]:
```
import sklearn
from sklearn.model_selection import train_test_split, KFold, LeaveOneGroupOut
from sklearn.linear_model import LogisticRegression, LinearRegression
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.metrics import roc_auc_score, accuracy_score, confusion_matrix, roc_curve, auc
from sklearn.preprocessing import OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.model_selection import cross_val_score
from mlxtend.feature_selection import SequentialFeatureSelector
```

In [3]:
```
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import Pipeline
```

Data Cleaning

In [4]:
```
# Loading in dataset
df = pd.read_csv("HR Dataset.csv")

# Dropping redundant features
df = df.drop(columns = ['CF_attrition label', 'CF_current Employee', 'Standard Hours', 'Employee Count', 'emp no', 'Employee N

## One hot encoding

df['Attrition'] = df['Attrition'].map({'Yes': 1, 'No': 0})
df['Gender'] = df['Gender'].map({'Female': 1, 'Male': 0})
df['Over Time'] = df['Over Time'].map({'Yes': 1, 'No': 0})

#Multi category encoding
colnames= ['Business Travel', 'CF_age band', 'Department', 'Education Field', 'Job Role', 'Marital Status', 'Education']
multi_df = df[colnames]
multi_encoded_df = pd.get_dummies(multi_df, dtype=int)

df = df.drop(columns = colnames)
df = pd.concat([df, multi_encoded_df], axis = 1)

df.head(2)
```

Out[4]:

	Attrition	Gender	Over Time	Training Times Last Year	Age	Distance From Home	Environment Satisfaction	Hourly Rate	Job Involvement	Job Level	...	Job Role_Sales Executive	Job Role_Sales Representative	Marita Status_Divorce
0	1	1	1	0	41	1	2	94	3	2	...	1	0	
1	0	0	0	3	49	8	3	61	2	2	...	0	0	

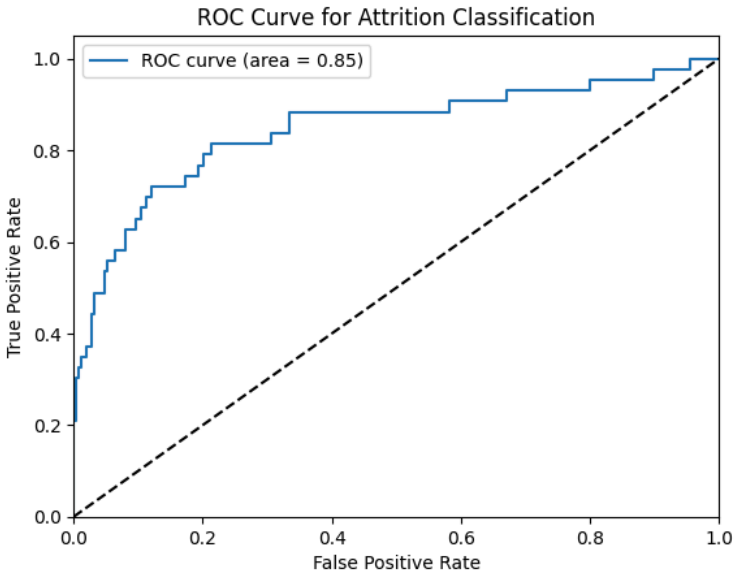
2 rows × 59 columns

Attrition Analysis and Prediction

In [5]:
```
x = df.drop(columns = ['Attrition'])
y = df['Attrition']

kf = KFold(n_splits = 10, shuffle = True, random_state = 0)

accuracy_list = []
roc_auc_list = []

for train_idx, test_idx in kf.split(x):

    model = LogisticRegression()

    x_train, x_test = x.iloc[train_idx], x.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    scaler = StandardScaler()
```

```
    x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

    model.fit(x_train, y_train)

    y_pred_prob = model.predict_proba(x_test)[:, 1]

    y_pred = (y_pred_prob > 0.5).astype(int)

    fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    roc_auc_list.append(roc_auc)

    accuracy = accuracy_score(y_test, y_pred)
    accuracy_list.append(accuracy)

avg_roc_auc = sum(roc_auc_list)/len(roc_auc_list)
avg_accuracy = sum(accuracy_list)/len(accuracy_list)

print('Average accuracy is', avg_accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % avg_roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Average accuracy is 0.8830835943709383
```



Oversampling

In [6]:
```
from imblearn.under_sampling import RandomUnderSampler

oversampler = SMOTE()
#under = RandomUnderSampler(sampling_strategy=0.5)
model = LogisticRegression()
scaler = StandardScaler()

x = df.drop(columns = ['Attrition'])
y = df['Attrition']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

x_train, y_train = oversampler.fit_resample(x_train, y_train)

model.fit(x_train, y_train)

y_pred_prob = model.predict_proba(x_test)[:, 1]

y_pred = (y_pred_prob > 0.5).astype(int)
```

```
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)
accuracy = accuracy_score(y_test, y_pred)

print('roc', roc_auc)
print('acc', accuracy)
```

```
roc 0.8276904964884504
acc 0.7675213675213676
```

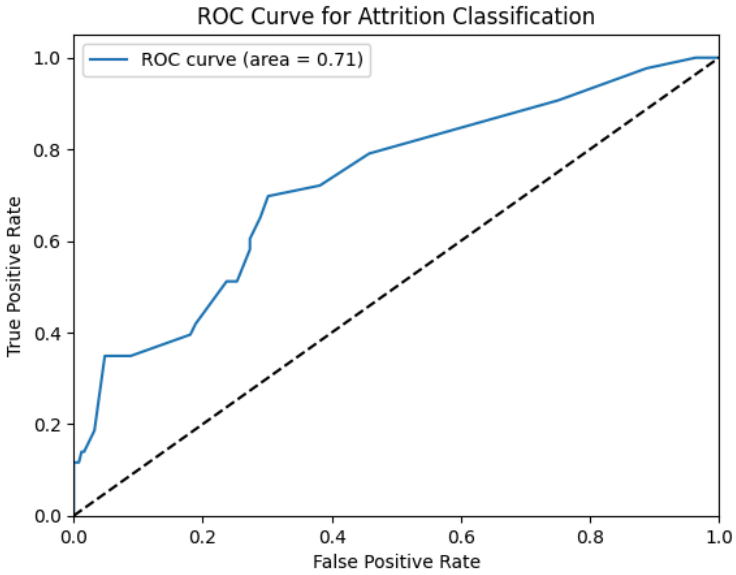
Feature selection

In [7]:
```
model = LogisticRegression()
forward = SequentialFeatureSelector(model, k_features=4, forward=True, verbose=1, scoring="neg_mean_squared_error")

x = df.iloc[:,1:]
y = df['Attrition']

scaler = StandardScaler()
x_scaled = scaler.fit_transform(x)

feature_selector = forward.fit(x_scaled,y)
feat_names = list(feature_selector.k_feature_names_)
print(feat_names)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  58 out of  58 | elapsed:    1.1s finished
Features: 1/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  57 out of  57 | elapsed:    1.1s finished
Features: 2/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  56 out of  56 | elapsed:    1.1s finished
Features: 3/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  55 out of  55 | elapsed:   25.3s finished
Features: 4/4['0', '1', '7', '29']
```

In [8]:
```
forward_df = x.iloc[:, [0, 1, 7, 29]]
# features selected by forward selection are gender, over time, job involvement, under 25

kf = KFold(n_splits = 10, shuffle = True, random_state = 0)

accuracy_list = []
roc_auc_list = []

for train_idx, test_idx in kf.split(forward_df):

    model = LogisticRegression()

    x_train, x_test = forward_df.iloc[train_idx], forward_df.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    scaler = StandardScaler()

    x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

    model.fit(x_train, y_train)

    y_pred_prob = model.predict_proba(x_test)[:, 1]

    y_pred = (y_pred_prob > 0.5).astype(int)

    fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    roc_auc_list.append(roc_auc)

    accuracy = accuracy_score(y_test, y_pred)
    accuracy_list.append(accuracy)

avg_roc_auc = sum(roc_auc_list)/len(roc_auc_list)
avg_accuracy = sum(accuracy_list)/len(accuracy_list)

print('Average accuracy is', avg_accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % avg_roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Average accuracy is 0.8396500537659545
```



In [9]:
```
model = LogisticRegression()
forward = SequentialFeatureSelector(model, k_features=4, forward=False, verbose=1, scoring="neg_mean_squared_error")

x = df.drop(columns = ['Attrition'])
y = df['Attrition']

scaler = StandardScaler()
x_scaled = scaler.fit_transform(x)

feature_selector = forward.fit(x_scaled,y)
feat_names = list(feature_selector.k_feature_names_)
print(feat_names)
```

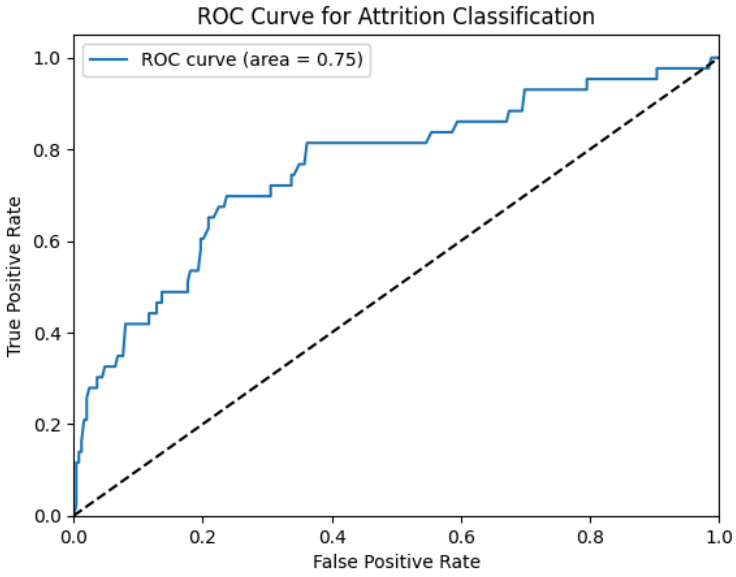
```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  58 out of  58 | elapsed:  2.0min finished
Features: 57/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  57 out of  57 | elapsed:  1.8min finished
Features: 56/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  56 out of  56 | elapsed:  2.0min finished
Features: 55/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  55 out of  55 | elapsed:  1.7min finished
Features: 54/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  54 out of  54 | elapsed:  1.7min finished
Features: 53/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  53 out of  53 | elapsed:  1.6min finished
Features: 52/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  52 out of  52 | elapsed:  1.7min finished
Features: 51/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  51 out of  51 | elapsed:  1.9min finished
Features: 50/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  50 out of  50 | elapsed:  1.8min finished
Features: 49/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  49 out of  49 | elapsed:  1.4min finished
Features: 48/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  48 out of  48 | elapsed:  1.4min finished
Features: 47/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  47 out of  47 | elapsed:  1.4min finished
Features: 46/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  46 out of  46 | elapsed:  1.3min finished
Features: 45/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  45 out of  45 | elapsed:  1.3min finished
Features: 44/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  44 out of  44 | elapsed:  1.3min finished
Features: 43/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  43 out of  43 | elapsed:  1.2min finished
Features: 42/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  42 out of  42 | elapsed:  1.1min finished
Features: 41/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  41 out of  41 | elapsed:  1.1min finished
Features: 40/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  40 out of  40 | elapsed:   57.1s finished
Features: 39/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  39 out of  39 | elapsed:  1.0min finished
Features: 38/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  38 out of  38 | elapsed:  1.1min finished
Features: 37/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  37 out of  37 | elapsed:   56.1s finished
Features: 36/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  36 out of  36 | elapsed:   53.5s finished
Features: 35/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  35 out of  35 | elapsed:   40.4s finished
Features: 34/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  34 out of  34 | elapsed:   39.5s finished
Features: 33/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  33 out of  33 | elapsed:   38.8s finished
Features: 32/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  32 out of  32 | elapsed:   36.8s finished
Features: 31/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  31 out of  31 | elapsed:   32.7s finished
Features: 30/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  30 out of  30 | elapsed:   35.2s finished
Features: 29/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  29 out of  29 | elapsed:   26.9s finished
Features: 28/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  28 out of  28 | elapsed:   27.4s finished
Features: 27/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  27 out of  27 | elapsed:   25.5s finished
Features: 26/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  26 out of  26 | elapsed:   20.1s finished
Features: 25/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  25 out of  25 | elapsed:   17.3s finished
Features: 24/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  24 out of  24 | elapsed:   15.7s finished
Features: 23/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  23 out of  23 | elapsed:   15.1s finished
Features: 22/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  22 out of  22 | elapsed:   13.0s finished
Features: 21/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  21 out of  21 | elapsed:   11.8s finished
Features: 20/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  20 out of  20 | elapsed:   11.6s finished
Features: 19/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  19 out of  19 | elapsed:   10.8s finished
Features: 18/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  18 out of  18 | elapsed:    8.6s finished
Features: 17/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  17 out of  17 | elapsed:    7.2s finished
Features: 16/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  16 out of  16 | elapsed:    8.2s finished
```

```
Features: 15/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  15 out of  15 | elapsed:    6.4s finished
Features: 14/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  14 out of  14 | elapsed:    8.1s finished
Features: 13/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  13 out of  13 | elapsed:    6.5s finished
Features: 12/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  12 out of  12 | elapsed:    6.0s finished
Features: 11/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  11 out of  11 | elapsed:    6.1s finished
Features: 10/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:    4.6s finished
Features: 9/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   9 out of   9 | elapsed:    2.7s finished
Features: 8/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   8 out of   8 | elapsed:    3.6s finished
Features: 7/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   7 out of   7 | elapsed:    2.0s finished
Features: 6/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   6 out of   6 | elapsed:    2.8s finished
Features: 5/4[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   5 out of   5 | elapsed:    2.3s finished
Features: 4/4['1', '4', '8', '15']
```

In [10]:

```
backward_df = x.iloc[:, [1, 4, 8, 15]]
# features selcted by backward selection are overtime, distance from home, job level, stock option level

kf = KFold(n_splits = 10, shuffle = True, random_state = 0)

accuracy_list = []
roc_auc_list = []

for train_idx, test_idx in kf.split(backward_df):

    model = LogisticRegression()

    x_train, x_test = backward_df.iloc[train_idx], backward_df.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    scaler = StandardScaler()

    x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

    model.fit(x_train, y_train)

    y_pred_prob = model.predict_proba(x_test)[:, 1]

    y_pred = (y_pred_prob > 0.5).astype(int)

    fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    roc_auc_list.append(roc_auc)

    accuracy = accuracy_score(y_test, y_pred)
    accuracy_list.append(accuracy)

avg_roc_auc = sum(roc_auc_list)/len(roc_auc_list)
avg_accuracy = sum(accuracy_list)/len(accuracy_list)

print('Average accuracy is', avg_accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % avg_roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Average accuracy is 0.851955444387302
```



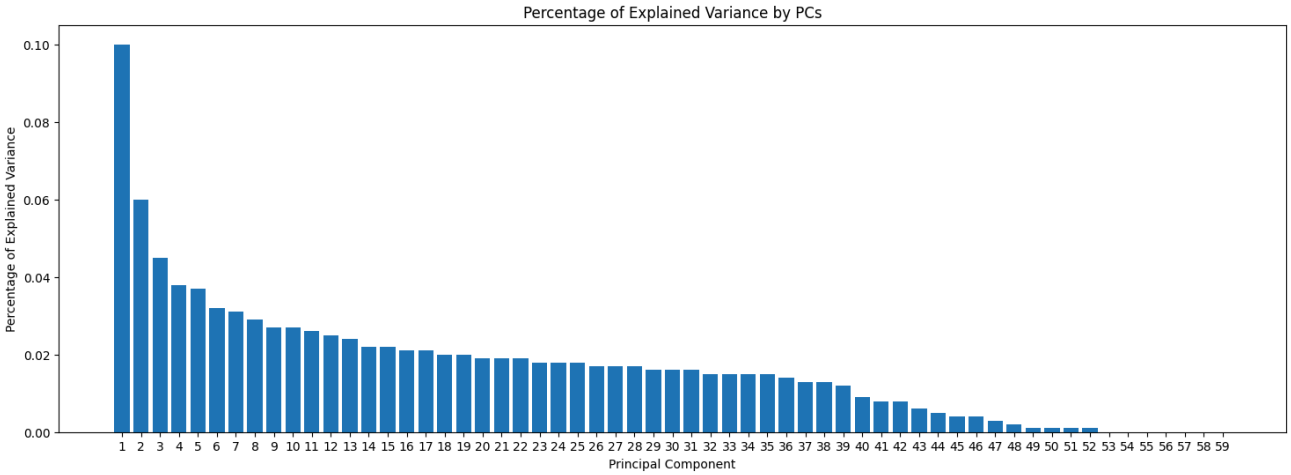
PCA Exploration

In [11]:
```
#PCA
scaler = StandardScaler()
scaled_df = scaler.fit_transform(df)
#pca = PCA(n_components=10)
#pca.fit_transform(scaled_df)

pca = PCA()
pca.fit(scaled_df)
pca_data = pca.transform(scaled_df)
per_var = np.round(pca.explained_variance_ratio_, decimals = 3)
labels = [str(x) for x in range(1, len(per_var)+1)]

print(labels)

'''
# Finding optimal number of PCs
nrange = np.arange(59)
var_list = []
for n in nrange:
    pca = PCA(n_components = n)
    pca.fit(scaled_df)
    var_list.append(np.sum(pca.explained_variance_ratio_))

plt.figure(figsize=(4,2),dpi=150)
plt.grid()
plt.plot(nrange,var_list)
plt.xlabel('Number of Components')
plt.ylabel('Explained variance ratio')
plt.title('Components vs. Explained Variance Ratio')
'''
```

```
['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '2
3', '24', '25', '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38', '39', '40', '41', '42', '43', '4
4', '45', '46', '47', '48', '49', '50', '51', '52', '53', '54', '55', '56', '57', '58', '59']
```

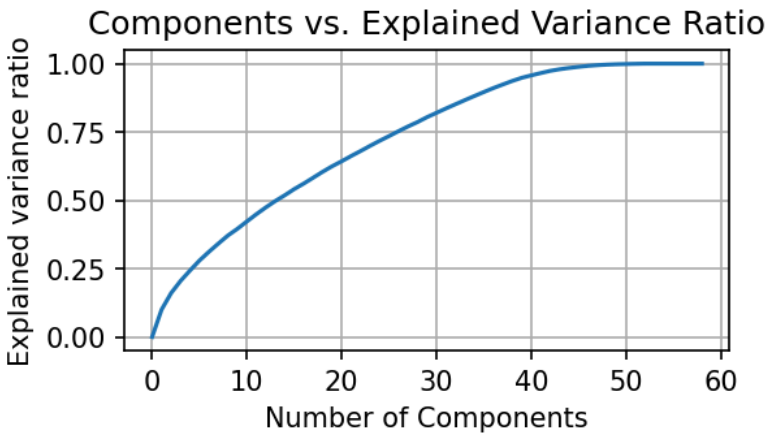
Out[11]:
```
"\n# Finding optimal number of PCs\nnrange = np.arange(59)\nvar_list = []\nfor n in nrange:\n    pca = PCA(n_components = n)
 \n    pca.fit(scaled_df)\n    var_list.append(np.sum(pca.explained_variance_ratio_))\n\nplt.figure(figsize=(4,2),dpi=150)\np
lt.grid()\nplt.plot(nrange,var_list)\nplt.xlabel('Number of Components')\nplt.ylabel('Explained variance ratio')\nplt.title
('Components vs. Explained Variance Ratio')\n"
```

In [12]:
```
plt.figure(figsize=(18,6))
plt.bar(x=range(1,len(per_var)+1), height=per_var, tick_label=labels)
plt.ylabel('Percentage of Explained Variance')
plt.xlabel('Principal Component')
plt.title('Percentage of Explained Variance by PCs')
plt.show()
```



In [13]:
```
# Finding optimal number of PCs
nrange = np.arange(59)
var_list = []
for n in nrange:
    pca = PCA(n_components = n)
    pca.fit(scaled_df)
    var_list.append(np.sum(pca.explained_variance_ratio_))

plt.figure(figsize=(4,2),dpi=150)
plt.grid()
plt.plot(nrange,var_list)
plt.xlabel('Number of Components')
plt.ylabel('Explained variance ratio')
plt.title('Components vs. Explained Variance Ratio')
```

Out[13]: Text(0.5, 1.0, 'Components vs. Explained Variance Ratio')



In [14]:
```
for count, variance in enumerate(var_list):
    if variance > 0.8:
        print('The number of PCs that explain 80% of variance is', count)
        break

for count, variance in enumerate(var_list):
    if variance > 0.90:
        print('The number of PCs that explain 90% of variance is',count)
        break
```

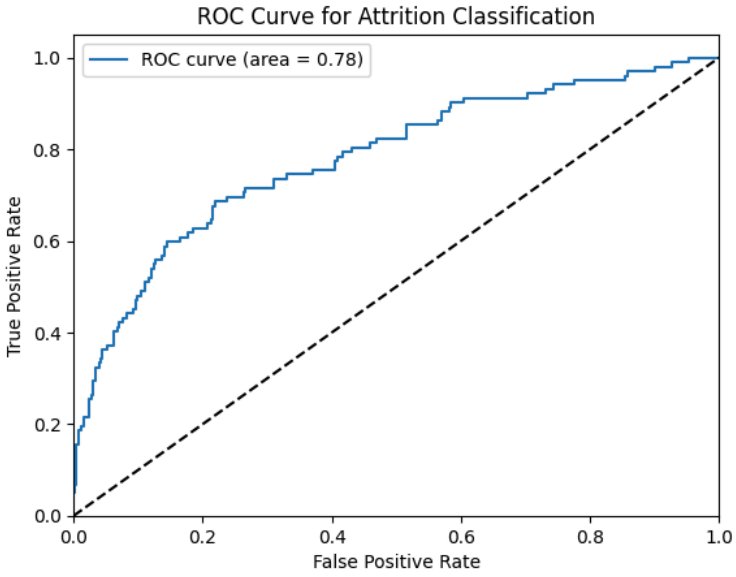
```
The number of PCs that explain 80% of variance is 29
The number of PCs that explain 90% of variance is 36
```

In [15]:
```
x = df.drop(columns = ['Attrition'])
y = df['Attrition']
```

In [16]:
```
# Dimensionaity reduction (explain 80% of variance = 29 )

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

scaler = StandardScaler()

x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
```

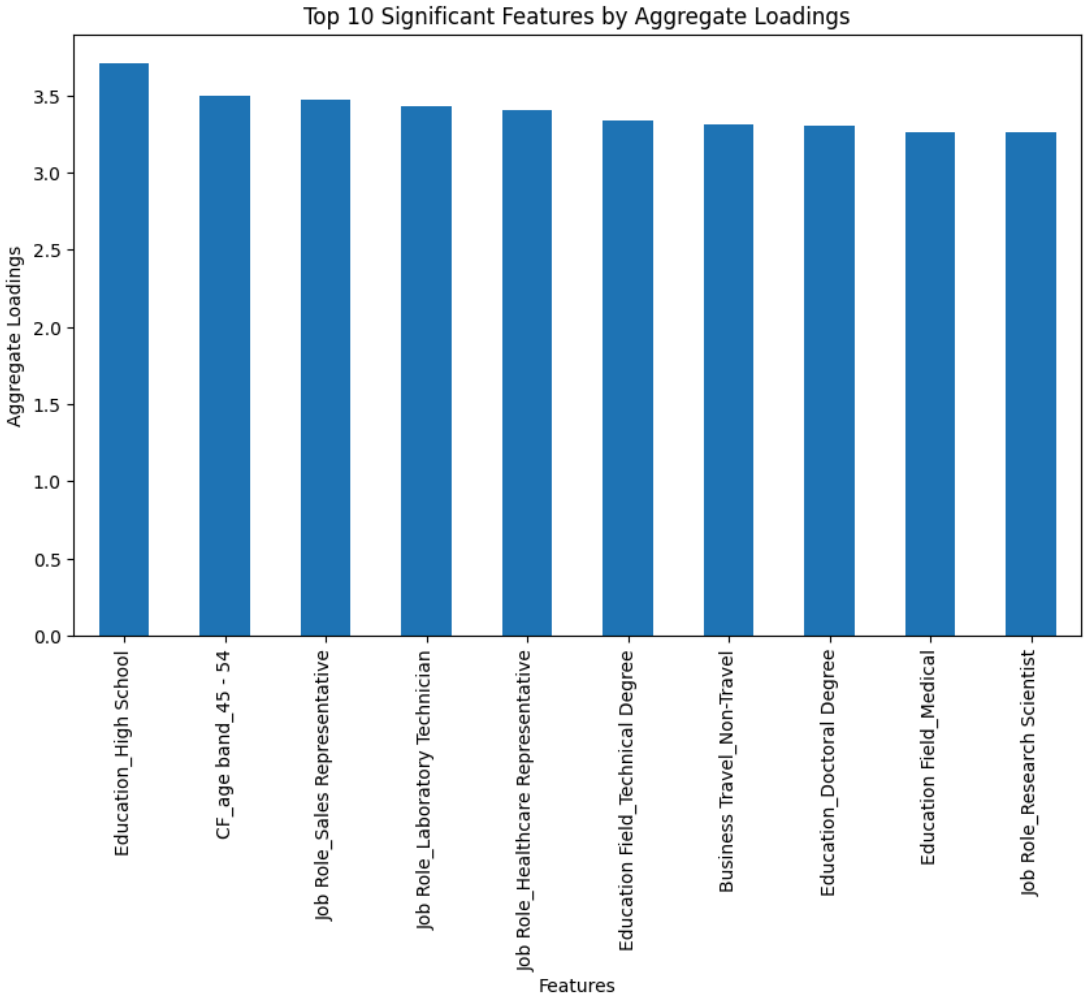
```
pca = PCA(n_components=29)
pca.fit(x_train)

x_train_pca = pca.transform(x_train)
x_test_pca = pca.transform(x_test)

model = LogisticRegression(solver = 'lbfgs')
model.fit(x_train_pca, y_train)

y_pred_prob = model.predict_proba(x_test_pca)[:, 1]
y_pred = (y_pred_prob > 0.5).astype(int)

#y_pred = model.predict(x_test_pca)
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)
accuracy = accuracy_score(y_test, y_pred)
print('Model accuracy for 29 components is', accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Model accuracy for 29 components is 0.852991452991453
```



In [17]:

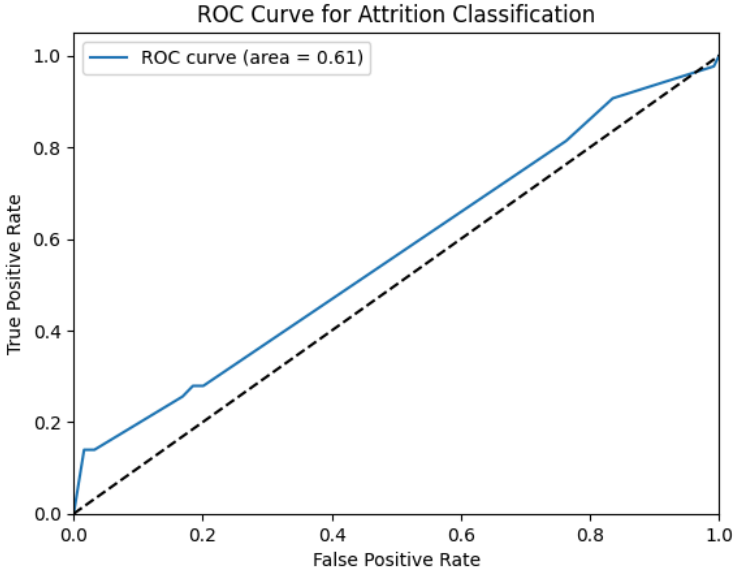
```
explained_variance = pca.explained_variance_ratio_
cumulative_explained_variance = np.cumsum(explained_variance)

loadings = pca.components_.T
loadings_df = pd.DataFrame(loadings, index=x.columns, columns=[f'PC{i+1}' for i in range(29)])
aggregate_loadings = loadings_df.abs().sum(axis=1).sort_values(ascending=False)

top_10_features = aggregate_loadings.head(10)

plt.figure(figsize=(10, 6))
top_10_features.plot(kind='bar')
plt.title(f'Top 10 Significant Features by Aggregate Loadings')
plt.xlabel('Features')
plt.ylabel('Aggregate Loadings')
plt.show()
```



In [18]:
```
refined_df = x[['Education_High School', 'CF_age band_45 - 54',
'Job Role_Sales Representative', 'Job Role_Laboratory Technician']]

kf = KFold(n_splits = 10, shuffle = True, random_state = 0)

accuracy_list = []
roc_auc_list = []

for train_idx, test_idx in kf.split(refined_df):

    model = LogisticRegression()

    x_train, x_test = refined_df.iloc[train_idx], refined_df.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    scaler = StandardScaler()

    x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

    model.fit(x_train, y_train)

    y_pred_prob = model.predict_proba(x_test)[:, 1]

    y_pred = (y_pred_prob > 0.5).astype(int)

    fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    roc_auc_list.append(roc_auc)

    accuracy = accuracy_score(y_test, y_pred)
    accuracy_list.append(accuracy)

avg_roc_auc = sum(roc_auc_list)/len(roc_auc_list)
avg_accuracy = sum(accuracy_list)/len(accuracy_list)

print('Average accuracy is', avg_accuracy)

plt.figure()
```

```
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % avg_roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

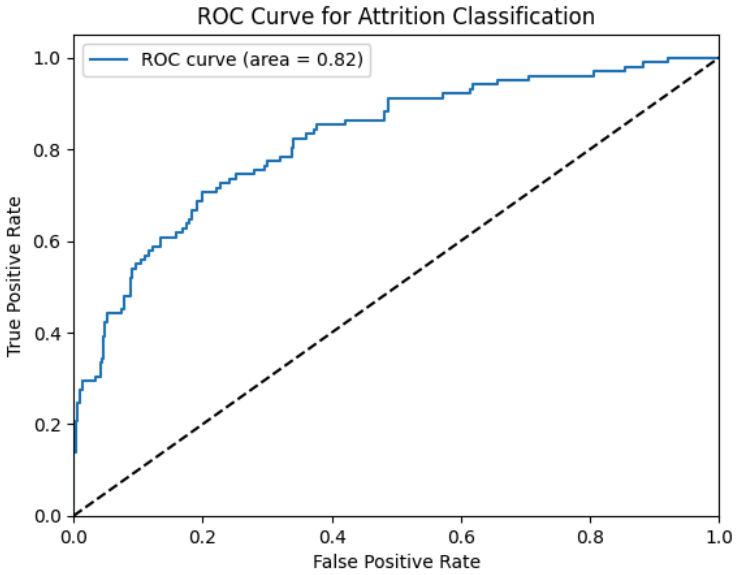
```
Average accuracy is 0.8317885361634485
```



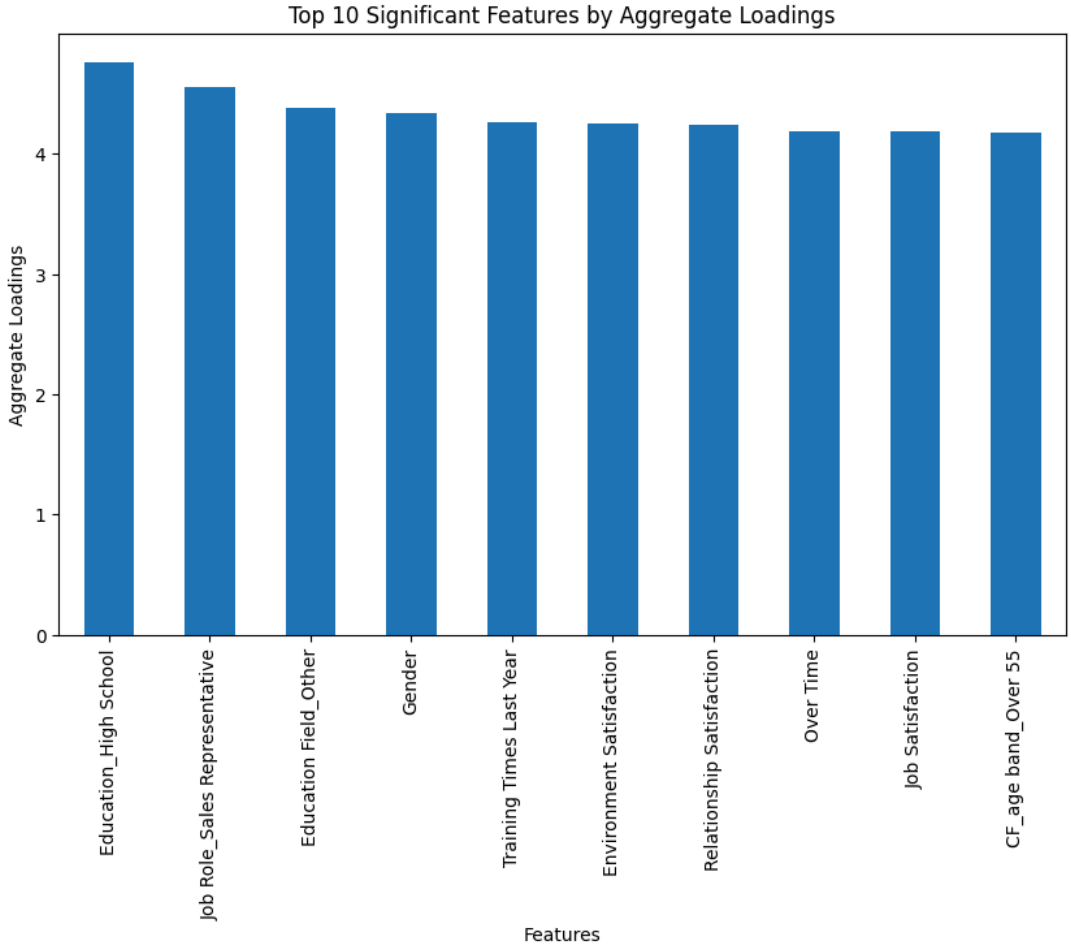
In [19]:

```
# Information retention (explain 90% of variance = 36 )

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

scaler = StandardScaler()

x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

pca = PCA(n_components=36)
pca.fit(x_train)

x_train_pca = pca.transform(x_train)
x_test_pca = pca.transform(x_test)

model = LogisticRegression(solver = 'lbfgs')
model.fit(x_train_pca, y_train)

y_pred_prob = model.predict_proba(x_test_pca)[:, 1]
y_pred = (y_pred_prob > 0.5).astype(int)

#y_pred = model.predict(x_test_pca)
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)
accuracy = accuracy_score(y_test, y_pred)
print('Model accuracy for 36 components is', accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Model accuracy for 36 components is 0.8495726495726496
```



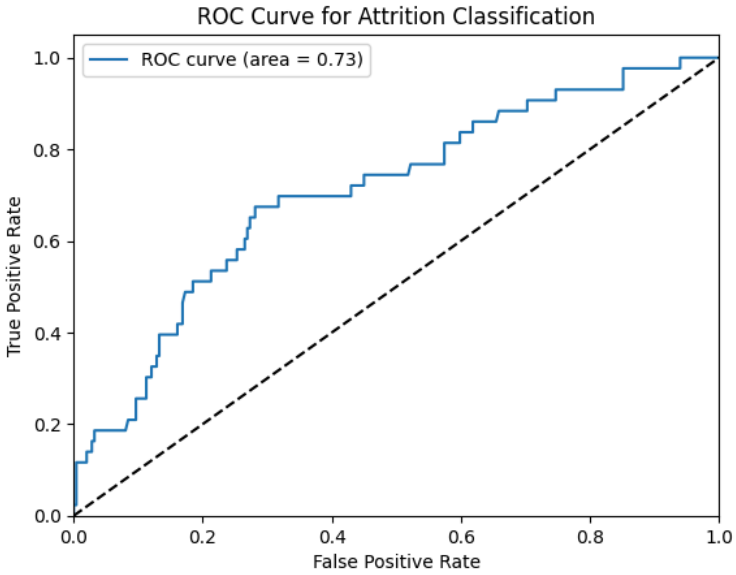
In [20]:
```
explained_variance = pca.explained_variance_ratio_
cumulative_explained_variance = np.cumsum(explained_variance)

loadings = pca.components_.T
loadings_df = pd.DataFrame(loadings, index=x.columns, columns=[f'PC{i+1}' for i in range(36)])
aggregate_loadings = loadings_df.abs().sum(axis=1).sort_values(ascending=False)

top_10_features = aggregate_loadings.head(10)

plt.figure(figsize=(10, 6))
top_10_features.plot(kind='bar')
plt.title(f'Top 10 Significant Features by Aggregate Loadings')
plt.xlabel('Features')
plt.ylabel('Aggregate Loadings')
plt.show()
```



In [21]:
```
refined_df = x[['Education_High School', 'Job Role_Sales Representative',
'Education Field_Other', 'Gender', 'Training Times Last Year',
'Environment Satisfaction', 'Relationship Satisfaction', 'Over Time',
'Job Satisfaction', 'CF_age band_Over 55']]

kf = KFold(n_splits = 10, shuffle = True, random_state = 0)

accuracy_list = []
roc_auc_list = []

for train_idx, test_idx in kf.split(refined_df):

    model = LogisticRegression()

    x_train, x_test = refined_df.iloc[train_idx], refined_df.iloc[test_idx]
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    scaler = StandardScaler()

    x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

    model.fit(x_train, y_train)

    y_pred_prob = model.predict_proba(x_test)[:, 1]

    y_pred = (y_pred_prob > 0.5).astype(int)

    fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    roc_auc_list.append(roc_auc)

    accuracy = accuracy_score(y_test, y_pred)
    accuracy_list.append(accuracy)

avg_roc_auc = sum(roc_auc_list)/len(roc_auc_list)
avg_accuracy = sum(accuracy_list)/len(accuracy_list)

print('Average accuracy is', avg_accuracy)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % avg_roc_auc)
```

```
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

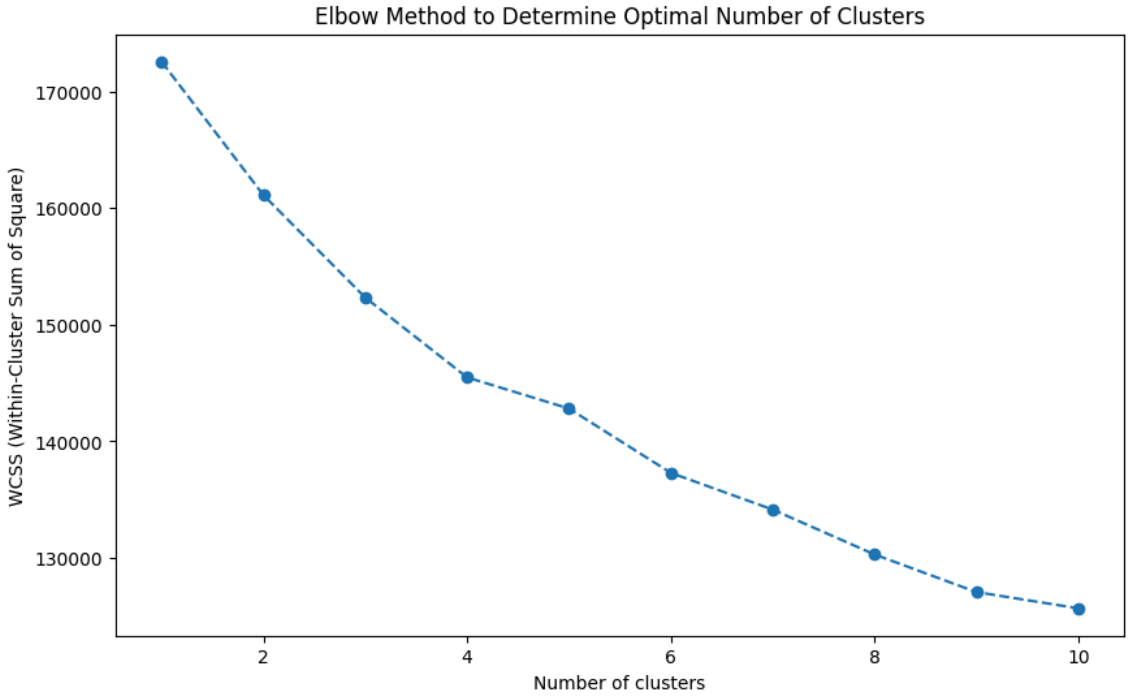
```
Average accuracy is 0.8379400626490252
```



Employee Segmentation for Targeted HR Policies

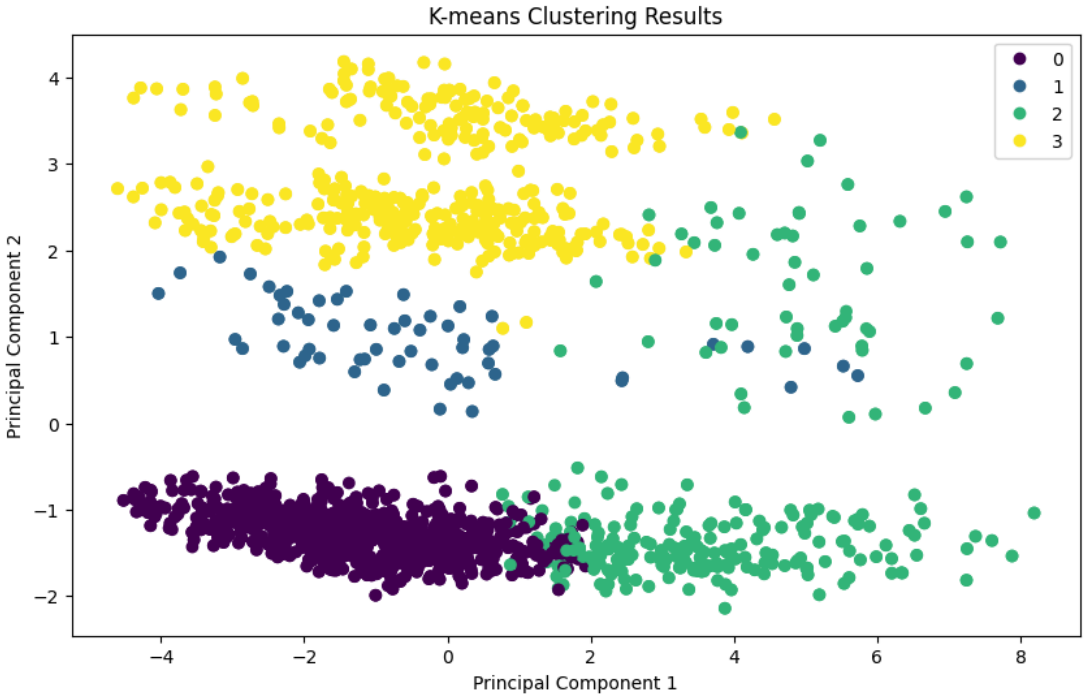
In [22]:
```
# Scaling the data after one-hot encoding in preparation for K-means

df_encoded = df.copy()

scaler = StandardScaler()
numerical_cols = df_encoded.select_dtypes(include=['int64', 'float64']).columns.tolist()

imputer = SimpleImputer(strategy='mean')
df_encoded[numerical_cols] = imputer.fit_transform(df_encoded[numerical_cols])

scaler = StandardScaler()
df_encoded[numerical_cols] = scaler.fit_transform(df_encoded[numerical_cols])
```

In [23]:
```
# Determine optimal number of clusters using Elbow method

wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(df_encoded)
    wcss.append(kmeans.inertia_)

plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss, marker='o', linestyle='--')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS (Within-Cluster Sum of Square)')
plt.title('Elbow Method to Determine Optimal Number of Clusters')
plt.show()
```



In [24]:
```
# K-means with optimal k

optimal_clusters = 4
kmeans = KMeans(n_clusters=optimal_clusters, random_state=42)
df_encoded['Cluster'] = kmeans.fit_predict(df_encoded)

print(df_encoded['Cluster'].value_counts())

pca = PCA(n_components=2)
principal_components = pca.fit_transform(df_encoded.drop(columns=['Cluster']))
df_pca = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])
df_pca['Cluster'] = df_encoded['Cluster']

plt.figure(figsize=(10, 6))
scatter = plt.scatter(df_pca['PC1'], df_pca['PC2'], c=df_pca['Cluster'], cmap='viridis')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-means Clustering Results')
plt.legend(handles=scatter.legend_elements()[0], labels=set(df_pca['Cluster']))
plt.show()
```

```
Cluster
0    1482
3     795
2     532
1     116
Name: count, dtype: int64
```



In [25]:
```python
# Gather summary statistics of each cluster

import seaborn as sns

print(df_encoded['Cluster'].value_counts())

def profile_clusters(df, clusters_col='Cluster'):
    profile = {}

    for cluster in df[clusters_col].unique():
        cluster_data = df[df[clusters_col] == cluster]

        numerical_summary = cluster_data.describe().transpose()

        categorical_summary = {}
        for col in df.select_dtypes(include=['object', 'category']).columns:
            categorical_summary[col] = cluster_data[col].mode()[0], cluster_data[col].value_counts().to_dict()

        profile[cluster] = {
            'numerical_summary': numerical_summary,
            'categorical_summary': categorical_summary
        }

    return profile

cluster_profiles = profile_clusters(df_encoded)

for cluster, profile in cluster_profiles.items():
    print(f"Cluster {cluster} Profile:")
    print("Numerical Summary:")
    print(profile['numerical_summary'])
    print("\nCategorical Summary:")
    for col, summary in profile['categorical_summary'].items():
        print(f"{col}: Mode - {summary[0]}, Value Counts - {summary[1]}")
    print("\n")

plt.figure(figsize=(12, 6))
sns.boxplot(x='Cluster', y='Age', data=df_encoded)
plt.title('Age Distribution Across Clusters')
plt.show()

plt.figure(figsize=(12, 6))
sns.countplot(x='Cluster', hue='Gender', data=df_encoded)
plt.title('Gender Distribution Across Clusters')
plt.show()
```

```
Cluster
0    1482
3     795
2     532
1     116
Name: count, dtype: int64
Cluster 3 Profile:
Numerical Summary:
                                  count      mean           std       min  \
Attrition                         795.0  0.138807  1.108392e+00 -0.449688
Gender                            795.0  0.012138  1.003023e+00 -0.817078
Over Time                         795.0  0.017390  1.008840e+00 -0.627841
Training Times Last Year          795.0  0.066646  9.708438e-01 -2.172710
Age                               795.0 -0.210154  8.937497e-01 -2.069965
Distance From Home                795.0  0.027960  9.782930e-01 -1.010282
Environment Satisfaction          795.0 -0.036901  9.818338e-01 -1.577180
Hourly Rate                       795.0 -0.031207  1.007729e+00 -1.765871
Job Involvement                   795.0 -0.050373  9.797328e-01 -2.432156
Job Level                         795.0 -0.037479  5.973129e-01 -0.963087
Job Satisfaction                  795.0  0.047855  1.003170e+00 -1.567069
Monthly Income                    795.0 -0.136678  5.478627e-01 -1.158907
Num Companies Worked              795.0 -0.107419  9.522780e-01 -1.080064
Percent Salary Hike               795.0 -0.033622  9.629397e-01 -1.152589
Performance Rating                795.0 -0.056854  9.427591e-01 -0.426401
Relationship Satisfaction         795.0 -0.047246  1.042395e+00 -1.583114
Stock Option Level                795.0 -0.016409  9.902336e-01 -0.932311
Total Working Years               795.0 -0.263202  6.944872e-01 -1.449367
Work Life Balance                 795.0  0.097744  9.117327e-01 -2.494650
Years At Company                  795.0 -0.148174  7.134420e-01 -1.143398
Years In Current Role             795.0 -0.041075  9.018841e-01 -1.166720
Years Since Last Promotion        795.0 -0.071008  8.655196e-01 -0.678761
Years With Curr Manager           795.0 -0.072092  9.215153e-01 -1.155012
Business Travel_Non-Travel        795.0  0.018498  1.024423e+00 -0.338062
Business Travel_Travel_Frequently 795.0  0.014528  1.012065e+00 -0.481227
Business Travel_Travel_Rarely     795.0 -0.024862  1.011732e+00 -1.562426
CF_age band_25 - 34               795.0  0.175524  1.029985e+00 -0.775232
CF_age band_35 - 44               795.0 -0.048083  9.834797e-01 -0.723490
CF_age band_45 - 54               795.0 -0.097449  9.042628e-01 -0.448039
CF_age band_Over 55               795.0 -0.138328  6.228951e-01 -0.221674
CF_age band_Under 25              795.0  0.006429  1.011757e+00 -0.266526
CF_age band_Under 26              795.0  0.049548  1.918464e+00 -0.018493
CF_age band_Under 27              795.0  0.049548  1.918464e+00 -0.018493
Department_HR                     795.0 -0.212170  5.554610e-17 -0.212170
Department_R&D                    795.0 -1.367648  2.221844e-16 -1.367648
Department_Sales                  795.0  1.508468  2.221844e-16  1.508468
Education Field_Human Resources   795.0 -0.137145  2.777305e-17 -0.137145
Education Field_Life Sciences     795.0 -0.158411  9.586135e-01 -0.835827
Education Field_Marketing         795.0  0.798351  1.540344e+00 -0.349255
Education Field_Medical           795.0 -0.256767  8.550901e-01 -0.678998
Education Field_Other             795.0 -0.078427  8.312234e-01 -0.242930
Education Field_Technical Degree  795.0 -0.051596  9.224163e-01 -0.314977
Job Role_Healthcare Representative 795.0 -0.313006  0.000000e+00 -0.313006
Job Role_Human Resources          795.0 -0.192006  2.777305e-17 -0.192006
Job Role_Laboratory Technician    795.0 -0.458998  1.110922e-16 -0.458998
Job Role_Manager                  795.0 -0.254058  2.779527e-01 -0.273811
Job Role_Manufacturing Director   795.0 -0.330477  5.554610e-17 -0.330477
Job Role_Research Director        795.0 -0.240554  5.554610e-17 -0.240554
Job Role_Research Scientist       795.0 -0.497328  5.554610e-17 -0.497328
Job Role_Sales Executive          795.0  1.348863  9.905002e-01 -0.535051
Job Role_Sales Representative     795.0  0.661376  1.755688e+00 -0.246852
Marital Status_Divorced           795.0 -0.057269  9.599411e-01 -0.535051
Marital Status_Married            795.0 -0.010616  9.996798e-01 -0.919470
Marital Status_Single             795.0  0.062443  1.022652e+00 -0.684917
Education_Associates Degree       795.0  0.023493  1.018692e+00 -0.485532
Education_Bachelor's Degree       795.0 -0.046723  9.889338e-01 -0.799734
Education_Doctoral Degree         795.0 -0.014772  9.609765e-01 -0.184213
Education_High School             795.0  0.002670  1.003848e+00 -0.360250
Education_Master's Degree         795.0  0.034493  1.017628e+00 -0.610406
Cluster                           795.0  3.000000  0.000000e+00  3.000000

                                      25%       50%       75%       max
Attrition                       -0.449688 -0.449688 -0.449688   2.223763
Gender                          -0.817078 -0.817078  1.223873   1.223873
Over Time                       -0.627841 -0.627841  1.592759   1.592759
Training Times Last Year        -0.620585  0.155478  0.155478   2.483666
Age                             -0.757352 -0.319814  0.336493   2.524181
Distance From Home              -0.824771 -0.144562  0.473810   2.452600
Environment Satisfaction        -0.661307  0.254566  1.170439   1.170439
Hourly Rate                     -0.929780  0.004674  0.840765   1.676856
Job Involvement                 -1.025382  0.381392  0.381392   1.788166
Job Level                       -0.060193 -0.060193 -0.060193   1.745595
Job Satisfaction                -0.661025  0.245019  1.151063   1.151063
Monthly Income                  -0.478222 -0.227632  0.214559   1.501666
Num Companies Worked            -0.678999 -0.678999  0.524195   2.529519
Percent Salary Hike             -0.879209 -0.332449  0.761071   2.674731
Performance Rating              -0.426401 -0.426401 -0.426401   2.345208
```

```
Relationship Satisfaction          -0.658102  0.266909  1.191921   1.191921
Stock Option Level                 -0.932311  0.242112  0.242112   2.590957
Total Working Years                -0.679361 -0.294357  0.090646   3.042336
Work Life Balance                   0.336330  0.336330  0.336330   1.751820
Years At Company                   -0.654397 -0.328397  0.323604   2.442608
Years In Current Role              -0.615001 -0.339142  0.764296   3.247030
Years Since Last Promotion         -0.678761 -0.368643 -0.058525   3.973013
Years With Curr Manager            -0.594699 -0.314542  0.806085   3.607653
Business Travel_Non-Travel         -0.338062 -0.338062 -0.338062   2.958040
Business Travel_Travel_Frequently  -0.481227 -0.481227 -0.481227   2.078024
Business Travel_Travel_Rarely      -1.562426  0.640030  0.640030   0.640030
CF_age band_25 - 34                -0.775232 -0.775232  1.289936   1.289936
CF_age band_35 - 44                -0.723490 -0.723490  1.382189   1.382189
CF_age band_45 - 54                -0.448039 -0.448039 -0.448039   2.231949
CF_age band_Over 55                -0.221674 -0.221674 -0.221674   4.511138
CF_age band_Under 25               -0.266526 -0.266526 -0.266526   3.751975
CF_age band_Under 26               -0.018493 -0.018493 -0.018493  54.074023
CF_age band_Under 27               -0.018493 -0.018493 -0.018493  54.074023
Department_HR                      -0.212170 -0.212170 -0.212170  -0.212170
Department_R&D                     -1.367648 -1.367648 -1.367648  -1.367648
Department_Sales                    1.508468  1.508468  1.508468   1.508468
Education Field_Human Resources    -0.137145 -0.137145 -0.137145  -0.137145
Education Field_Life Sciences      -0.835827 -0.835827  1.196420   1.196420
Education Field_Marketing          -0.349255 -0.349255  2.863235   2.863235
Education Field_Medical            -0.678998 -0.678998 -0.678998   1.472757
Education Field_Other              -0.242930 -0.242930 -0.242930   4.116404
Education Field_Technical Degree   -0.314977 -0.314977 -0.314977   3.174830
Job Role_Healthcare Representative -0.313006 -0.313006 -0.313006  -0.313006
Job Role_Human Resources           -0.192006 -0.192006 -0.192006  -0.192006
Job Role_Laboratory Technician     -0.458998 -0.458998 -0.458998  -0.458998
Job Role_Manager                   -0.273811 -0.273811 -0.273811   3.652155
Job Role_Manufacturing Director    -0.330477 -0.330477 -0.330477  -0.330477
Job Role_Research Director         -0.240554 -0.240554 -0.240554  -0.240554
Job Role_Research Scientist        -0.497328 -0.497328 -0.497328  -0.497328
Job Role_Sales Executive            1.868980  1.868980  1.868980   1.868980
Job Role_Sales Representative      -0.246852 -0.246852 -0.246852   4.051014
Marital Status_Divorced            -0.535051 -0.535051 -0.535051   1.868980
Marital Status_Married             -0.919470 -0.919470  1.087583   1.087583
Marital Status_Single              -0.684917 -0.684917  1.460031   1.460031
Education_Associates Degree        -0.485532 -0.485532 -0.485532   2.059596
Education_Bachelor's Degree        -0.799734 -0.799734  1.250416   1.250416
Education_Doctoral Degree          -0.184213 -0.184213 -0.184213   5.428513
Education_High School              -0.360250 -0.360250 -0.360250   2.775853
Education_Master's Degree          -0.610406 -0.610406  1.638255   1.638255
Cluster                             3.000000  3.000000  3.000000   3.000000

Categorical Summary:


Cluster 0 Profile:
Numerical Summary:
                                   count      mean           std       min  \
Attrition                         1482.0  0.021142  1.018706e+00 -0.449688
Gender                            1482.0 -0.040360  9.912698e-01 -0.817078
Over Time                         1482.0 -0.016502  9.922040e-01 -0.627841
Training Times Last Year          1482.0 -0.013664  1.034301e+00 -2.172710
Age                               1482.0 -0.255969  8.843423e-01 -2.069965
Distance From Home                1482.0  0.004148  1.002993e+00 -1.010282
Environment Satisfaction          1482.0  0.017255  1.005348e+00 -1.577180
Hourly Rate                       1482.0  0.019276  9.953239e-01 -1.765871
Job Involvement                   1482.0  0.027325  1.020350e+00 -2.432156
Job Level                         1482.0 -0.525043  5.542986e-01 -0.963087
Job Satisfaction                  1482.0  0.009032  1.004806e+00 -1.567069
Monthly Income                    1482.0 -0.512769  4.231617e-01 -1.168031
Num Companies Worked              1482.0 -0.050881  9.869276e-01 -1.080064
Percent Salary Hike               1482.0  0.024679  9.979962e-01 -1.152589
Performance Rating                1482.0  0.018702  1.017956e+00 -0.426401
Relationship Satisfaction         1482.0 -0.021455  9.962728e-01 -1.583114
Stock Option Level                1482.0  0.005959  1.014021e+00 -0.932311
Total Working Years               1482.0 -0.389786  5.938747e-01 -1.449367
Work Life Balance                 1482.0 -0.063866  1.030235e+00 -2.494650
Years At Company                  1482.0 -0.293861  6.153886e-01 -1.143398
Years In Current Role             1482.0 -0.235462  8.335376e-01 -1.166720
Years Since Last Promotion        1482.0 -0.217141  6.912558e-01 -0.678761
Years With Curr Manager           1482.0 -0.202630  8.566621e-01 -1.155012
Business Travel_Non-Travel        1482.0  0.017793  1.023233e+00 -0.338062
Business Travel_Travel_Frequently 1482.0  0.012664  1.010322e+00 -0.481227
Business Travel_Travel_Rarely     1482.0 -0.022787  1.010539e+00 -1.562426
CF_age band_25 - 34               1482.0  0.136117  1.025789e+00 -0.775232
CF_age band_35 - 44               1482.0  0.043762  1.013708e+00 -0.723490
CF_age band_45 - 54               1482.0 -0.236461  7.229203e-01 -0.448039
CF_age band_Over 55               1482.0 -0.103513  7.386732e-01 -0.221674
CF_age band_Under 25              1482.0  0.096820  1.152816e+00 -0.266526
CF_age band_Under 26              1482.0 -0.018493  0.000000e+00 -0.018493
CF_age band_Under 27              1482.0 -0.018493  0.000000e+00 -0.018493
Department_HR                     1482.0 -0.212170  5.552989e-17 -0.212170
```

Department_R&D	1482.0	0.731182	3.331793e-16	0.731182	
Department_Sales	1482.0	-0.662924	0.000000e+00	-0.662924	
Education Field_Human Resources	1482.0	-0.137145	0.000000e+00	-0.137145	
Education Field_Life Sciences	1482.0	0.115846	1.014420e+00	-0.835827	
Education Field_Marketing	1482.0	-0.349255	5.552989e-17	-0.349255	
Education Field_Medical	1482.0	0.109398	1.037108e+00	-0.678998	
Education Field_Other	1482.0	0.060046	1.108969e+00	-0.242930	
Education Field_Technical Degree	1482.0	0.019404	1.027534e+00	-0.314977	
Job Role_Healthcare Representative	1482.0	0.110679	1.143498e+00	-0.313006	
Job Role_Human Resources	1482.0	-0.192006	2.776494e-17	-0.192006	
Job Role_Laboratory Technician	1482.0	0.436240	1.249373e+00	-0.458998	
Job Role_Manager	1482.0	-0.268513	1.441753e-01	-0.273811	
Job Role_Manufacturing Director	1482.0	0.172305	1.198215e+00	-0.330477	
Job Role_Research Director	1482.0	-0.234619	1.614963e-01	-0.240554	
Job Role_Research Scientist	1482.0	0.474086	1.222187e+00	-0.497328	
Job Role_Sales Executive	1482.0	-0.535051	1.110598e-16	-0.535051	
Job Role_Sales Representative	1482.0	-0.246852	5.552989e-17	-0.246852	
Marital Status_Divorced	1482.0	-0.014340	9.906201e-01	-0.535051	
Marital Status_Married	1482.0	-0.025641	9.978496e-01	-0.919470	
Marital Status_Single	1482.0	0.040197	1.015006e+00	-0.684917	
Education_Associates Degree	1482.0	0.010785	1.008735e+00	-0.485532	
Education_Bachelor's Degree	1482.0	0.020603	1.004760e+00	-0.799734	
Education_Doctoral Degree	1482.0	-0.047871	8.643856e-01	-0.184213	
Education_High School	1482.0	0.029118	1.034510e+00	-0.360250	
Education_Master's Degree	1482.0	-0.033826	9.822111e-01	-0.610406	
Cluster	1482.0	0.000000	0.000000e+00	0.000000	

	25%	50%	75%	max
Attrition	-0.449688	-0.449688	-0.449688	2.223763
Gender	-0.817078	-0.817078	1.223873	1.223873
Over Time	-0.627841	-0.627841	1.592759	1.592759
Training Times Last Year	-0.620585	0.155478	0.155478	2.483666
Age	-0.866736	-0.319814	0.227108	2.414797
Distance From Home	-0.886608	-0.268236	0.597485	2.452600
Environment Satisfaction	-0.661307	0.254566	1.170439	1.170439
Hourly Rate	-0.880598	0.053856	0.889947	1.676856
Job Involvement	-1.025382	0.381392	0.381392	1.788166
Job Level	-0.963087	-0.963087	-0.060193	0.842701
Job Satisfaction	-0.661025	0.245019	1.151063	1.151063
Monthly Income	-0.840101	-0.622613	-0.290227	1.519489
Num Companies Worked	-0.678999	-0.678999	0.524195	2.529519
Percent Salary Hike	-0.879209	-0.332449	0.761071	2.674731
Performance Rating	-0.426401	-0.426401	-0.426401	2.345208
Relationship Satisfaction	-0.658102	0.266909	1.191921	1.191921
Stock Option Level	-0.932311	0.242112	0.242112	2.590957
Total Working Years	-0.807695	-0.422692	-0.166023	1.758993
Work Life Balance	-1.079160	0.336330	0.336330	1.751820
Years At Company	-0.817397	-0.328397	0.160604	2.116607
Years In Current Role	-0.615001	-0.615001	0.212577	2.971171
Years Since Last Promotion	-0.678761	-0.368643	-0.058525	3.973013
Years With Curr Manager	-0.874855	-0.594699	0.806085	3.607653
Business Travel_Non-Travel	-0.338062	-0.338062	-0.338062	2.958040
Business Travel_Travel_Frequently	-0.481227	-0.481227	-0.481227	2.078024
Business Travel_Travel_Rarely	-1.562426	0.640030	0.640030	0.640030
CF_age band_25 - 34	-0.775232	-0.775232	1.289936	1.289936
CF_age band_35 - 44	-0.723490	-0.723490	1.382189	1.382189
CF_age band_45 - 54	-0.448039	-0.448039	-0.448039	2.231949
CF_age band_Over 55	-0.221674	-0.221674	-0.221674	4.511138
CF_age band_Under 25	-0.266526	-0.266526	-0.266526	3.751975
CF_age band_Under 26	-0.018493	-0.018493	-0.018493	-0.018493
CF_age band_Under 27	-0.018493	-0.018493	-0.018493	-0.018493
Department_HR	-0.212170	-0.212170	-0.212170	-0.212170
Department_R&D	0.731182	0.731182	0.731182	0.731182
Department_Sales	-0.662924	-0.662924	-0.662924	-0.662924
Education Field_Human Resources	-0.137145	-0.137145	-0.137145	-0.137145
Education Field_Life Sciences	-0.835827	-0.835827	1.196420	1.196420
Education Field_Marketing	-0.349255	-0.349255	-0.349255	-0.349255
Education Field_Medical	-0.678998	-0.678998	1.472757	1.472757
Education Field_Other	-0.242930	-0.242930	-0.242930	4.116404
Education Field_Technical Degree	-0.314977	-0.314977	-0.314977	3.174830
Job Role_Healthcare Representative	-0.313006	-0.313006	-0.313006	3.194823
Job Role_Human Resources	-0.192006	-0.192006	-0.192006	-0.192006
Job Role_Laboratory Technician	-0.458998	-0.458998	2.178661	2.178661
Job Role_Manager	-0.273811	-0.273811	-0.273811	3.652155
Job Role_Manufacturing Director	-0.330477	-0.330477	-0.330477	3.025930
Job Role_Research Director	-0.240554	-0.240554	-0.240554	4.157072
Job Role_Research Scientist	-0.497328	-0.497328	2.010747	2.010747
Job Role_Sales Executive	-0.535051	-0.535051	-0.535051	-0.535051
Job Role_Sales Representative	-0.246852	-0.246852	-0.246852	-0.246852
Marital Status_Divorced	-0.535051	-0.535051	-0.535051	1.868980
Marital Status_Married	-0.919470	-0.919470	1.087583	1.087583
Marital Status_Single	-0.684917	-0.684917	1.460031	1.460031
Education_Associates Degree	-0.485532	-0.485532	-0.485532	2.059596
Education_Bachelor's Degree	-0.799734	-0.799734	1.250416	1.250416
Education_Doctoral Degree	-0.184213	-0.184213	-0.184213	5.428513
Education_High School	-0.360250	-0.360250	-0.360250	2.775853

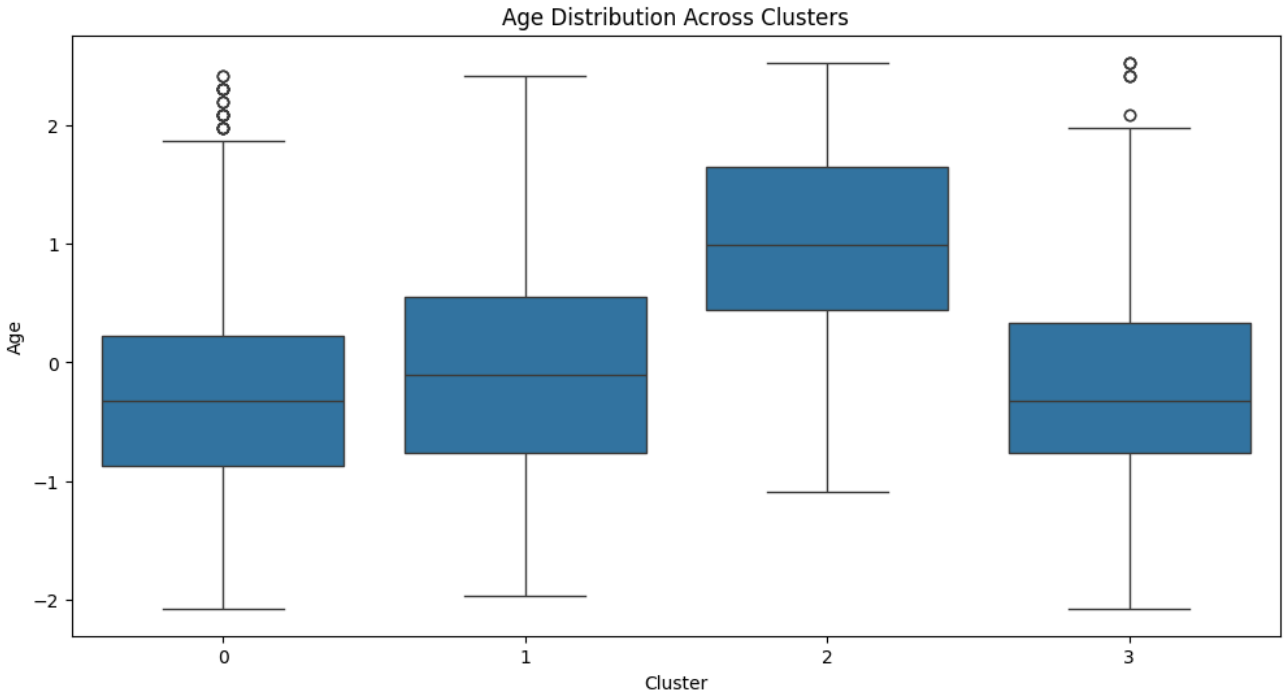
```
Education_Master's Degree          -0.610406 -0.610406  1.638255  1.638255
Cluster                             0.000000  0.000000  0.000000  0.000000

Categorical Summary:


Cluster 2 Profile:
Numerical Summary:
                                  count      mean           std       min  \
Attrition                         532.0 -0.288879  6.362520e-01 -0.449688
Gender                            532.0  0.126670  1.018545e+00 -0.817078
Over Time                         532.0  0.031660  1.015621e+00 -0.627841
Training Times Last Year          532.0 -0.022491  9.513868e-01 -2.172710
Age                               532.0  1.028575  7.885693e-01 -1.085505
Distance From Home                532.0 -0.038555  1.026168e+00 -1.010282
Environment Satisfaction          532.0  0.020433  1.031266e+00 -1.577180
Hourly Rate                       532.0  0.027971  9.994873e-01 -1.765871
Job Involvement                   532.0  0.000611  9.633775e-01 -2.432156
Job Level                         532.0  1.569089  7.872311e-01 -0.060193
Job Satisfaction                  532.0 -0.071756  9.829507e-01 -1.567069
Monthly Income                    532.0  1.669849  8.720953e-01 -0.883917
Num Companies Worked              532.0  0.276922  1.018106e+00 -1.080064
Percent Salary Hike               532.0  0.002596  1.064625e+00 -1.152589
Performance Rating                532.0  0.042480  1.040067e+00 -0.426401
Relationship Satisfaction         532.0  0.089558  9.553849e-01 -1.583114
Stock Option Level                532.0  0.008110  9.681851e-01 -0.932311
Total Working Years               532.0  1.515833  8.869980e-01 -0.294357
Work Life Balance                 532.0 -0.014882  1.011680e+00 -2.494650
Years At Company                  532.0  1.069974  1.457879e+00 -1.143398
Years In Current Role             532.0  0.767407  1.211286e+00 -1.166720
Years Since Last Promotion        532.0  0.757576  1.495702e+00 -0.678761
Years With Curr Manager           532.0  0.713402  1.192315e+00 -1.155012
Business Travel_Non-Travel        532.0 -0.077843  8.896521e-01 -0.338062
Business Travel_Travel_Frequently 532.0 -0.048271  9.603776e-01 -0.481227
Business Travel_Travel_Rarely     532.0  0.093556  9.521848e-01 -1.562426
CF_age band_25 - 34               532.0 -0.643248  5.055993e-01 -0.775232
CF_age band_35 - 44               532.0 -0.082287  9.699459e-01 -0.723490
CF_age band_45 - 54               532.0  0.841579  1.340307e+00 -0.448039
CF_age band_Over 55               532.0  0.490027  1.693286e+00 -0.221674
CF_age band_Under 25              532.0 -0.266526  5.556340e-17 -0.266526
CF_age band_Under 26              532.0 -0.018493  0.000000e+00 -0.018493
CF_age band_Under 27              532.0 -0.018493  0.000000e+00 -0.018493
Department_HR                     532.0 -0.119588  6.695320e-01 -0.212170
Department_R&D                    532.0  0.305104  8.450240e-01 -1.367648
Department_Sales                  532.0 -0.262931  8.425450e-01 -0.662924
Education Field_Human Resources   532.0 -0.137145  2.778170e-17 -0.137145
Education Field_Life Sciences     532.0 -0.003064  1.000383e+00 -0.835827
Education Field_Marketing         532.0 -0.143946  7.864894e-01 -0.349255
Education Field_Medical           532.0  0.138022  1.045256e+00 -0.678998
Education Field_Other             532.0 -0.046269  9.056380e-01 -0.242930
Education Field_Technical Degree  532.0  0.039251  1.054895e+00 -0.314977
Job Role_Healthcare Representative 532.0  0.227674  1.267794e+00 -0.313006
Job Role_Human Resources          532.0 -0.192006  2.778170e-17 -0.192006
Job Role_Laboratory Technician    532.0 -0.429250  2.787943e-01 -0.458998
Job Role_Manager                  532.0  1.098801  1.873864e+00 -0.273811
Job Role_Manufacturing Director   532.0  0.085919  1.107481e+00 -0.330477
Job Role_Research Director        532.0  1.065508  2.011313e+00 -0.240554
Job Role_Research Scientist       532.0 -0.469041  2.650976e-01 -0.497328
Job Role_Sales Executive          532.0 -0.408523  5.373180e-01 -0.535051
Job Role_Sales Representative     532.0 -0.246852  2.778170e-17 -0.246852
Marital Status_Divorced           532.0  0.106626  1.064421e+00 -0.535051
Marital Status_Married            532.0  0.046330  1.003761e+00 -0.919470
Marital Status_Single             532.0 -0.144648  9.319823e-01 -0.684917
Education_Associates Degree       532.0 -0.074102  9.378281e-01 -0.485532
Education_Bachelor's Degree       532.0 -0.005879  9.995970e-01 -0.799734
Education_Doctoral Degree         532.0  0.132295  1.295935e+00 -0.184213
Education_High School             532.0 -0.065503  9.160004e-01 -0.360250
Education_Master's Degree         532.0  0.065883  1.032172e+00 -0.610406
Cluster                           532.0  2.000000  0.000000e+00  2.000000

                                      25%       50%       75%       max
Attrition                        -0.449688 -0.449688 -0.449688  2.223763
Gender                           -0.817078 -0.817078  1.223873  1.223873
Over Time                        -0.627841 -0.627841  1.592759  1.592759
Training Times Last Year         -0.620585  0.155478  0.155478  2.483666
Age                               0.445877  0.992799  1.649106  2.524181
Distance From Home               -0.886608 -0.391910  0.473810  2.452600
Environment Satisfaction         -0.661307  0.254566  1.170439  1.170439
Hourly Rate                      -0.782235 -0.019917  0.889947  1.676856
Job Involvement                  -1.025382  0.381392  0.381392  1.788166
Job Level                         0.842701  1.745595  1.745595  2.648489
Job Satisfaction                 -0.661025  0.245019  1.151063  1.151063
Monthly Income                    0.958475  1.653271  2.410662  2.861340
Num Companies Worked             -0.678999  0.123130  0.925260  2.529519
Percent Salary Hike              -0.879209 -0.332449  0.761071  2.674731
Performance Rating               -0.426401 -0.426401 -0.426401  2.345208
```

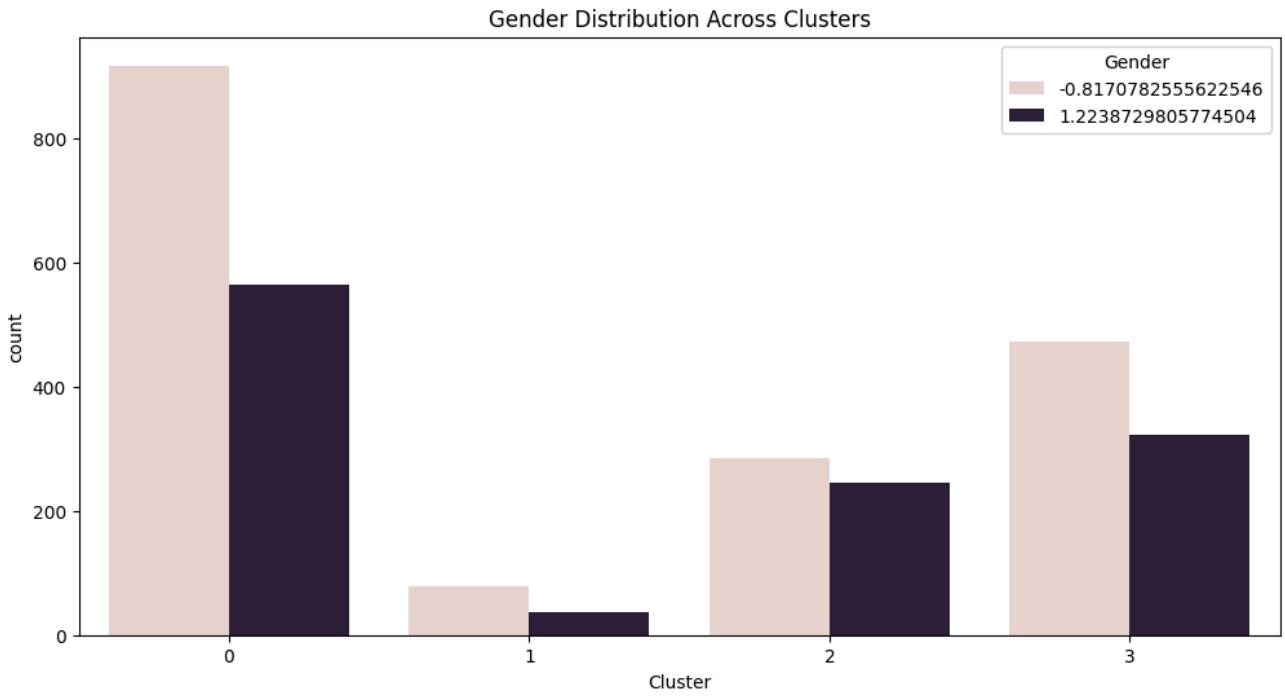
```
Relationship Satisfaction          -0.658102  0.266909  1.191921  1.191921
Stock Option Level                 -0.932311  0.242112  0.242112  2.590957
Total Working Years                 0.988986  1.502324  2.143996  3.684008
Work Life Balance                  -1.079160  0.336330  0.336330  1.751820
Years At Company                   -0.165396  0.812605  2.116607  5.376612
Years In Current Role              -0.339142  0.764296  1.316014  3.798749
Years Since Last Promotion         -0.368643  0.251593  1.492067  3.973013
Years With Curr Manager            -0.314542  0.806085  1.366399  3.607653
Business Travel_Non-Travel         -0.338062 -0.338062 -0.338062  2.958040
Business Travel_Travel_Frequently  -0.481227 -0.481227 -0.481227  2.078024
Business Travel_Travel_Rarely       0.640030  0.640030  0.640030  0.640030
CF_age band_25 - 34                -0.775232 -0.775232 -0.775232  1.289936
CF_age band_35 - 44                -0.723490 -0.723490  1.382189  1.382189
CF_age band_45 - 54                -0.448039 -0.448039  2.231949  2.231949
CF_age band_Over 55                -0.221674 -0.221674 -0.221674  4.511138
CF_age band_Under 25               -0.266526 -0.266526 -0.266526 -0.266526
CF_age band_Under 26               -0.018493 -0.018493 -0.018493 -0.018493
CF_age band_Under 27               -0.018493 -0.018493 -0.018493 -0.018493
Department_HR                      -0.212170 -0.212170 -0.212170  4.713203
Department_R&D                      0.731182  0.731182  0.731182  0.731182
Department_Sales                   -0.662924 -0.662924 -0.662924  1.508468
Education Field_Human Resources    -0.137145 -0.137145 -0.137145 -0.137145
Education Field_Life Sciences      -0.835827 -0.835827  1.196420  1.196420
Education Field_Marketing          -0.349255 -0.349255 -0.349255  2.863235
Education Field_Medical            -0.678998 -0.678998  1.472757  1.472757
Education Field_Other              -0.242930 -0.242930 -0.242930  4.116404
Education Field_Technical Degree   -0.314977 -0.314977 -0.314977  3.174830
Job Role_Healthcare Representative -0.313006 -0.313006 -0.313006  3.194823
Job Role_Human Resources           -0.192006 -0.192006 -0.192006 -0.192006
Job Role_Laboratory Technician     -0.458998 -0.458998 -0.458998  2.178661
Job Role_Manager                   -0.273811 -0.273811  3.652155  3.652155
Job Role_Manufacturing Director    -0.330477 -0.330477 -0.330477  3.025930
Job Role_Research Director         -0.240554 -0.240554  4.157072  4.157072
Job Role_Research Scientist        -0.497328 -0.497328 -0.497328  2.010747
Job Role_Sales Executive           -0.535051 -0.535051 -0.535051  1.868980
Job Role_Sales Representative      -0.246852 -0.246852 -0.246852 -0.246852
Marital Status_Divorced            -0.535051 -0.535051  1.868980  1.868980
Marital Status_Married             -0.919470 -0.919470  1.087583  1.087583
Marital Status_Single              -0.684917 -0.684917  1.460031  1.460031
Education_Associates Degree        -0.485532 -0.485532 -0.485532  2.059596
Education_Bachelor's Degree        -0.799734 -0.799734  1.250416  1.250416
Education_Doctoral Degree          -0.184213 -0.184213 -0.184213  5.428513
Education_High School              -0.360250 -0.360250 -0.360250  2.775853
Education_Master's Degree          -0.610406 -0.610406  1.638255  1.638255
Cluster                             2.000000  2.000000  2.000000  2.000000

Categorical Summary:


Cluster 1 Profile:
Numerical Summary:
                                  count      mean           std       min  \
Attrition                         116.0  0.103440  1.087662e+00 -0.449688
Gender                            116.0 -0.148491  9.620415e-01 -0.817078
Over Time                         116.0 -0.053548  9.765676e-01 -0.627841
Training Times Last Year          116.0 -0.179032  9.503362e-01 -2.172710
Age                               116.0 -0.006748  9.658305e-01 -1.960580
Distance From Home                116.0 -0.067798  9.960477e-01 -1.010282
Environment Satisfaction          116.0 -0.061253  9.127093e-01 -1.577180
Hourly Rate                       116.0 -0.160678  1.005658e+00 -1.765871
Job Involvement                   116.0 -0.006684  1.042652e+00 -2.432156
Job Level                         116.0 -0.231431  1.071061e+00 -0.963087
Job Satisfaction                  116.0 -0.114274  9.844286e-01 -1.567069
Monthly Income                    116.0 -0.170496  1.048043e+00 -1.052179
Num Companies Worked              116.0  0.116215  1.167179e+00 -1.080064
Percent Salary Hike               116.0 -0.096776  9.733426e-01 -1.152589
Performance Rating                116.0 -0.044110  9.598737e-01 -0.426401
Relationship Satisfaction         116.0  0.187167  9.174924e-01 -1.583114
Stock Option Level                116.0 -0.000872  1.044909e+00 -0.932311
Total Working Years               116.0 -0.168236  9.137493e-01 -1.321032
Work Life Balance                 116.0  0.214305  1.065307e+00 -2.494650
Years At Company                  116.0 -0.137293  8.061013e-01 -0.980398
Years In Current Role             116.0 -0.229749  7.699001e-01 -1.166720
Years Since Last Promotion        116.0 -0.213584  6.600812e-01 -0.678761
Years With Curr Manager           116.0 -0.188954  7.752890e-01 -1.155012
Business Travel_Non-Travel        116.0  0.002914  1.008161e+00 -0.338062
Business Travel_Travel_Frequently 116.0 -0.039976  9.709244e-01 -0.481227
Business Travel_Travel_Rarely     116.0  0.032456  9.886534e-01 -1.562426
CF_age band_25 - 34               116.0  0.008107  1.006399e+00 -0.775232
CF_age band_35 - 44               116.0  0.147826  1.041572e+00 -0.723490
CF_age band_45 - 54               116.0 -0.170799  8.197137e-01 -0.448039
CF_age band_Over 55               116.0  0.023127  1.052721e+00 -0.221674
CF_age band_Under 25              116.0 -0.058673  8.938368e-01 -0.266526
CF_age band_Under 26              116.0 -0.018493  6.968998e-18 -0.018493
CF_age band_Under 27              116.0 -0.018493  6.968998e-18 -0.018493
Department_HR                     116.0  4.713203  0.000000e+00  4.713203
```

Department_R&D	116.0	-1.367648	2.230079e-16	-1.367648
Department_Sales	116.0	-0.662924	0.000000e+00	-0.662924
Education Field_Human Resources	116.0	3.321039	3.721579e+00	-0.137145
Education Field_Life Sciences	116.0	-0.380323	8.511508e-01	-0.835827
Education Field_Marketing	116.0	-0.349255	1.672559e-16	-0.349255
Education Field_Medical	116.0	-0.270907	8.472086e-01	-0.678998
Education Field_Other	116.0	-0.017448	9.696484e-01	-0.242930
Education Field_Technical Degree	116.0	-0.074301	8.881379e-01	-0.314977
Job Role_Healthcare Representative	116.0	-0.313006	1.115040e-16	-0.313006
Job Role_Human Resources	116.0	4.649528	1.651722e+00	-0.192006
Job Role_Laboratory Technician	116.0	-0.458998	2.230079e-16	-0.458998
Job Role_Manager	116.0	0.132323	1.200814e+00	-0.273811
Job Role_Manufacturing Director	116.0	-0.330477	1.115040e-16	-0.330477
Job Role_Research Director	116.0	-0.240554	8.362797e-17	-0.240554
Job Role_Research Scientist	116.0	-0.497328	1.672559e-16	-0.497328
Job Role_Sales Executive	116.0	-0.535051	0.000000e+00	-0.535051
Job Role_Sales Representative	116.0	-0.246852	1.115040e-16	-0.246852
Marital Status_Divorced	116.0	0.086681	1.057236e+00	-0.535051
Marital Status_Married	116.0	0.187869	1.002473e+00	-0.919470
Marital Status_Single	116.0	-0.278116	8.445282e-01	-0.684917
Education_Associates Degree	116.0	0.041046	1.035455e+00	-0.485532
Education_Bachelor's Degree	116.0	0.083952	1.019682e+00	-0.799734
Education_Doctoral Degree	116.0	0.106101	1.248440e+00	-0.184213
Education_High School	116.0	-0.089896	8.840264e-01	-0.360250
Education_Master's Degree	116.0	-0.106396	9.417896e-01	-0.610406
Cluster	116.0	1.000000	0.000000e+00	1.000000

	25%	50%	75%	max
Attrition	-0.449688	-0.449688	-0.449688	2.223763
Gender	-0.817078	-0.817078	1.223873	1.223873
Over Time	-0.627841	-0.627841	1.592759	1.592759
Training Times Last Year	-0.620585	-0.232553	0.155478	2.483666
Age	-0.757352	-0.101045	0.555262	2.414797
Distance From Home	-0.886608	-0.391910	0.473810	2.081577
Environment Satisfaction	-0.661307	0.254566	0.254566	1.170439
Hourly Rate	-0.978962	-0.364189	0.693220	1.676856
Job Involvement	-1.025382	0.381392	0.381392	1.788166
Job Level	-0.963087	-0.963087	-0.060193	2.648489
Job Satisfaction	-0.661025	0.245019	1.151063	1.151063
Monthly Income	-0.838085	-0.624735	-0.026482	2.784317
Num Companies Worked	-0.678999	-0.277934	0.524195	2.529519
Percent Salary Hike	-0.879209	-0.332449	0.487691	2.127971
Performance Rating	-0.426401	-0.426401	-0.426401	2.345208
Relationship Satisfaction	-0.658102	0.266909	1.191921	1.191921
Stock Option Level	-0.932311	0.242112	0.242112	2.590957
Total Working Years	-0.679361	-0.422692	0.090646	3.042336
Work Life Balance	0.336330	0.336330	0.336330	1.751820
Years At Company	-0.654397	-0.328397	0.323604	2.442608
Years In Current Role	-0.615001	-0.615001	0.488436	1.591874
Years Since Last Promotion	-0.678761	-0.368643	-0.058525	3.042658
Years With Curr Manager	-0.594699	-0.454620	0.525928	1.646555
Business Travel_Non-Travel	-0.338062	-0.338062	-0.338062	2.958040
Business Travel_Travel_Frequently	-0.481227	-0.481227	-0.481227	2.078024
Business Travel_Travel_Rarely	-1.562426	0.640030	0.640030	0.640030
CF_age band_25 - 34	-0.775232	-0.775232	1.289936	1.289936
CF_age band_35 - 44	-0.723490	-0.723490	1.382189	1.382189
CF_age band_45 - 54	-0.448039	-0.448039	-0.448039	2.231949
CF_age band_Over 55	-0.221674	-0.221674	-0.221674	4.511138
CF_age band_Under 25	-0.266526	-0.266526	-0.266526	3.751975
CF_age band_Under 26	-0.018493	-0.018493	-0.018493	-0.018493
CF_age band_Under 27	-0.018493	-0.018493	-0.018493	-0.018493
Department_HR	4.713203	4.713203	4.713203	4.713203
Department_R&D	-1.367648	-1.367648	-1.367648	-1.367648
Department_Sales	-0.662924	-0.662924	-0.662924	-0.662924
Education Field_Human Resources	-0.137145	-0.137145	7.291548	7.291548
Education Field_Life Sciences	-0.835827	-0.835827	-0.835827	1.196420
Education Field_Marketing	-0.349255	-0.349255	-0.349255	-0.349255
Education Field_Medical	-0.678998	-0.678998	-0.678998	1.472757
Education Field_Other	-0.242930	-0.242930	-0.242930	4.116404
Education Field_Technical Degree	-0.314977	-0.314977	-0.314977	3.174830
Job Role_Healthcare Representative	-0.313006	-0.313006	-0.313006	-0.313006
Job Role_Human Resources	5.208167	5.208167	5.208167	5.208167
Job Role_Laboratory Technician	-0.458998	-0.458998	-0.458998	-0.458998
Job Role_Manager	-0.273811	-0.273811	-0.273811	3.652155
Job Role_Manufacturing Director	-0.330477	-0.330477	-0.330477	-0.330477
Job Role_Research Director	-0.240554	-0.240554	-0.240554	-0.240554
Job Role_Research Scientist	-0.497328	-0.497328	-0.497328	-0.497328
Job Role_Sales Executive	-0.535051	-0.535051	-0.535051	-0.535051
Job Role_Sales Representative	-0.246852	-0.246852	-0.246852	-0.246852
Marital Status_Divorced	-0.535051	-0.535051	1.868980	1.868980
Marital Status_Married	-0.919470	1.087583	1.087583	1.087583
Marital Status_Single	-0.684917	-0.684917	-0.684917	1.460031
Education_Associates Degree	-0.485532	-0.485532	-0.485532	2.059596
Education_Bachelor's Degree	-0.799734	-0.799734	1.250416	1.250416
Education_Doctoral Degree	-0.184213	-0.184213	-0.184213	5.428513
Education_High School	-0.360250	-0.360250	-0.360250	2.775853

```
Education_Master's Degree          -0.610406 -0.610406 -0.610406  1.638255
Cluster                             1.000000  1.000000  1.000000  1.000000

Categorical Summary:
```





Job Satisfaction and Salary Analysis

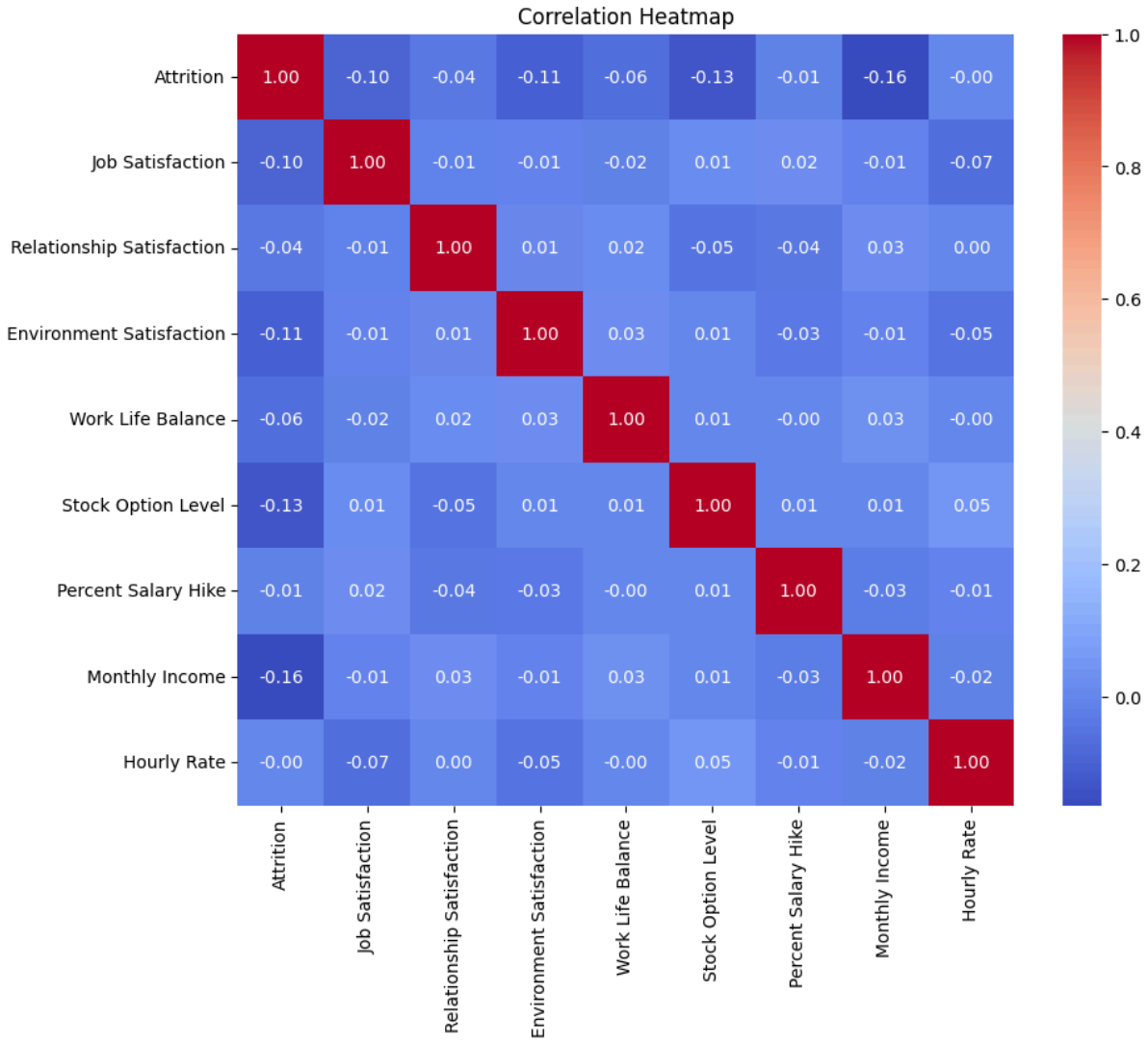
In [26]:
```
# Variables of Interest
# Satisfaction Variables: 'Job Satisfaction', 'Relationship Satisfaction', 'Environment Satisfaction', 'Work Life Balance'
# Salary Variables: 'Standard Hours', 'Stock Option Level', 'Percent Salary Hike', 'Monthly Income', 'Hourly Rate'
vars_of_interest = ['Attrition', 'Job Satisfaction', 'Relationship Satisfaction', 'Environment Satisfaction',
                    'Work Life Balance', 'Stock Option Level', 'Percent Salary Hike',
                    'Monthly Income', 'Hourly Rate']
obj3_subset = df[vars_of_interest]
```

In [27]:
```
obj3_subset.head(10)
```

Out[27]:

	Attrition	Job Satisfaction	Relationship Satisfaction	Environment Satisfaction	Work Life Balance	Stock Option Level	Percent Salary Hike	Monthly Income	Hourly Rate
0	1	4	1	2	1	0	11	5993	94
1	0	2	4	3	3	1	23	5130	61
2	1	3	2	4	3	0	15	2090	92
3	1	3	3	4	3	0	11	2909	56
4	1	2	4	1	3	1	12	3468	40
5	1	4	3	4	2	0	13	3068	79
6	1	1	1	3	2	3	20	2670	81
7	1	3	2	4	3	1	22	2693	67
8	1	3	2	4	3	0	21	9526	44
9	1	3	2	3	2	2	13	5237	94

In [28]:
```
# Initial Visualization
import seaborn as sns
plt.figure(figsize=(10, 8))
corr_matrix = obj3_subset.corr()
sns.heatmap(corr_matrix, cmap='coolwarm', annot=True, fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



In [29]:
```
# Creating Training and Test Set
from sklearn.model_selection import train_test_split
obj3_response = obj3_subset['Attrition']
```

```
obj3_predictors = obj3_subset[vars_of_interest[1:]]
x_train, x_test, y_train, y_test = train_test_split(obj3_predictors, obj3_response, test_size=0.2, random_state=1)
```

In [30]:
```
import statsmodels.api as sm
x_train = sm.add_constant(x_train)
x_test = sm.add_constant(x_test)
linear_model = sm.OLS(y_train, x_train)
linear_model = linear_model.fit()
linear_model.summary()
```

Out[30]:

OLS Regression Results

Dep. Variable:	Attrition	**R-squared:**	0.078
Model:	OLS	**Adj. R-squared:**	0.075
Method:	Least Squares	**F-statistic:**	24.80
Date:	Mon, 10 Jun 2024	**Prob (F-statistic):**	5.89e-37
Time:	19:26:02	**Log-Likelihood:**	-929.32
No. Observations:	2340	**AIC:**	1877.
Df Residuals:	2331	**BIC:**	1928.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6864	0.062	11.150	0.000	0.566	0.807
Job Satisfaction	-0.0366	0.007	-5.439	0.000	-0.050	-0.023
Relationship Satisfaction	-0.0180	0.007	-2.583	0.010	-0.032	-0.004
Environment Satisfaction	-0.0364	0.007	-5.356	0.000	-0.050	-0.023
Work Life Balance	-0.0319	0.011	-3.019	0.003	-0.053	-0.011
Stock Option Level	-0.0613	0.009	-6.887	0.000	-0.079	-0.044
Percent Salary Hike	-0.0021	0.002	-0.999	0.318	-0.006	0.002
Monthly Income	-1.381e-05	1.59e-06	-8.698	0.000	-1.69e-05	-1.07e-05
Hourly Rate	-0.0002	0.000	-0.575	0.565	-0.001	0.001

Omnibus:	559.172	**Durbin-Watson:**	2.082
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1033.447
Skew:	1.562	**Prob(JB):**	3.89e-225
Kurtosis:	3.918	**Cond. No.**	6.66e+04

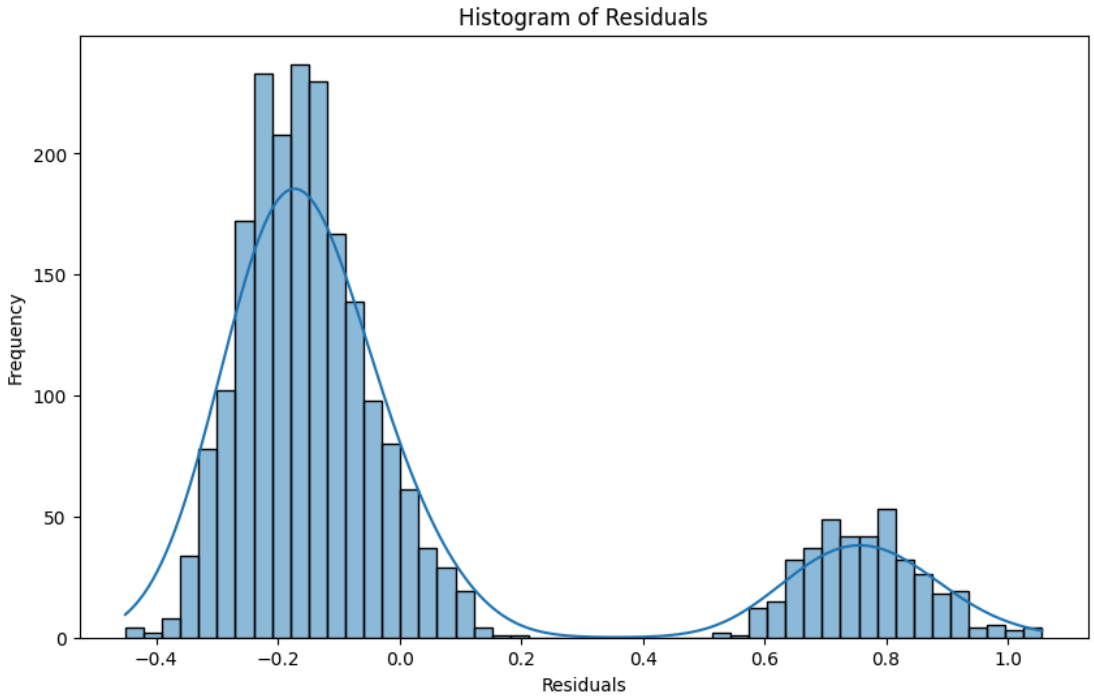
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.66e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

In [31]:
```
from sklearn.metrics import mean_squared_error
y_pred = linear_model.predict(x_test)
y_train_pred = linear_model.predict(x_train)
print(f'test error: {mean_squared_error(y_test, y_pred)}')
print(f'training error: {mean_squared_error(y_train, y_train_pred)}')
```

```
test error: 0.13210422263481308
training error: 0.12956296410564225
```

Test error and training error are similar signifying that the model is likely not overfitting. However, the R-squared value of the model is quite small likely due to the binary nature of the dependent variable.

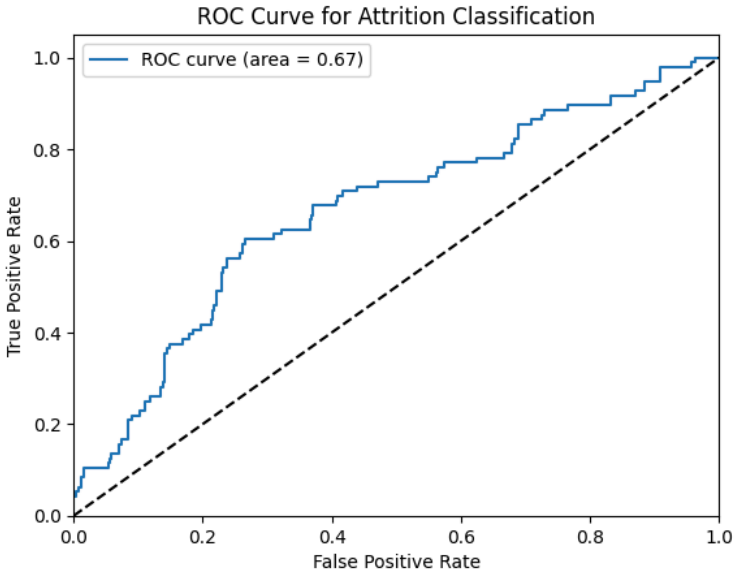
In [32]:
```
residuals = y_train - y_train_pred
plt.figure(figsize=(10, 6))
sns.histplot(residuals, kde=True)
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Histogram of Residuals')
plt.show()
```



Residual Plot would suggest that the relationship between these variables is not linear meaning the linear model is likely not an appropriate model for this subset of variables.

In [33]:
```
x_train, x_test, y_train, y_test = train_test_split(obj3_predictors, obj3_response, test_size=0.2, random_state=1)

scaler = StandardScaler()

x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)

x_const = sm.add_constant(x_train)

logistic_model = sm.Logit(y_train, x_const)
logistic_model = logistic_model.fit()

x_test_const = sm.add_constant(x_test)
y_pred_prob = logistic_model.predict(x_test_const)

y_pred = (y_pred_prob > 0.5).astype(int)

fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob)
roc_auc = auc(fpr, tpr)

plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve for Attrition Classification')
plt.legend()
plt.show()
```

```
Optimization terminated successfully.
         Current function value: 0.410665
         Iterations 7
```



In [34]:
```
print(accuracy_score(y_test, y_pred))
```

```
0.8376068376068376
```

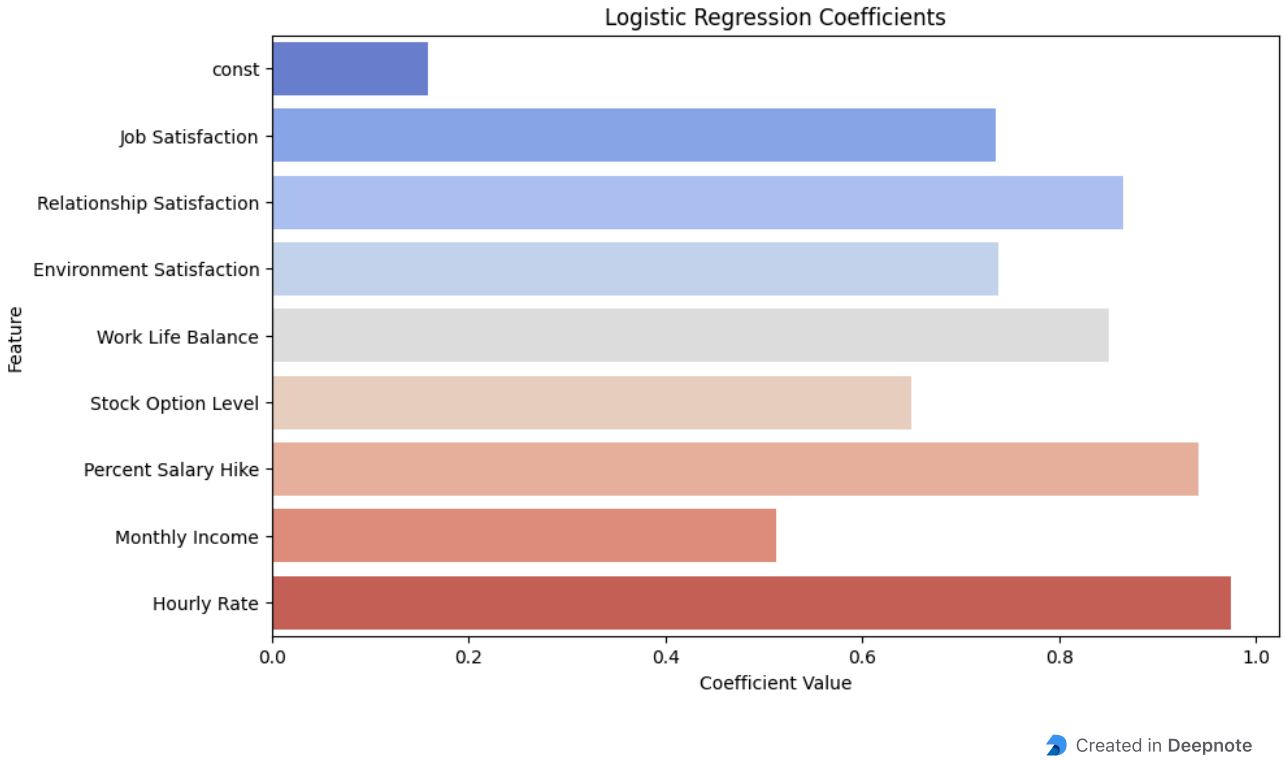
In [35]:
```
coefficients = logistic_model.params
odds_ratios = np.exp(coefficients)
conf_intervals = logistic_model.conf_int()
predictor_names = ['const','Job Satisfaction', 'Relationship Satisfaction', 'Environment Satisfaction',
                   'Work Life Balance', 'Stock Option Level', 'Percent Salary Hike',
                   'Monthly Income', 'Hourly Rate']
results_df = pd.DataFrame({
    'Predictor': predictor_names,
    'Odds Ratio': odds_ratios
})
results_df['Lower CI'] = conf_intervals[0]
results_df['Upper CI'] = conf_intervals[1]
results_df.iloc[1:9]
```

Out[35]:

	Predictor	Odds Ratio	Lower CI	Upper CI
x1	Job Satisfaction	0.735804	-0.419534	-0.194049
x2	Relationship Satisfaction	0.865428	-0.256089	-0.032972
x3	Environment Satisfaction	0.738466	-0.415738	-0.190621
x4	Work Life Balance	0.850452	-0.271671	-0.052305
x5	Stock Option Level	0.650045	-0.561957	-0.299470
x6	Percent Salary Hike	0.940857	-0.176333	0.054405
x7	Monthly Income	0.511778	-0.832948	-0.506780
x8	Hourly Rate	0.974691	-0.139703	0.088433

In [36]:
```
plt.figure(figsize=(10, 6))
sns.barplot(x = results_df['Odds Ratio'], y = results_df['Predictor'], palette="coolwarm")
plt.xlabel('Coefficient Value')
plt.ylabel('Feature')
plt.title('Logistic Regression Coefficients')
plt.show()
```

```
/tmp/ipykernel_99/1702464388.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and s
et `legend=False` for the same effect.

  sns.barplot(x = results_df['Odds Ratio'], y = results_df['Predictor'], palette="coolwarm")
```


Logistic Regression Coefficients
const
Job Satisfaction
Relationship Satisfaction
Environment Satisfaction
Work Life Balance
Stock Option Level
Percent Salary Hike
Monthly Income
Hourly Rate
Feature
0.0
0.2
0.4
0.6
0.8
1.0
Coefficient Value
Created in Deepnote