Springboard Data Science Career Track - Capstone Project

Performance Analysis of Online Training Services using Data Analytics and Machine Learning

Project Objectives

This project had two main objectives, a learning objective and a business objective.

Learning Objective:

- This capstone project summarizes the knowledge and skills of program participants in the data science career track as demonstrated in the project workflow from data gathering, wrangling, analysis, summarizing, and recommendation for action plans.

Business Objective:

- Through the use of real business data, findings and conclusions from this project are useful for Online Training Service Providers to improve their successes after conducting more extensive studies

Methodology

The following methodology was adapted to assess performance of Online Training Services:

- **Dataset:** Identify where to find the required data and data retrieval
- **Data Wrangling:** Once the data is identified and retrieved, basic data preprocessing and quality control are conducted
- **Data Exploration:** When the data is ready for exploration, apply statistical methods to identify data clusters, correlations, and trends
- **Machine Learning for Prediction:** With basic understanding of the data, more advanced data analytics of machine learning is applied to the data to build predictive models
- **Summary:** conclusions and recommendations

Dataset

The dataset chosen was the Open University Learning Analytics Dataset available at: https://analyse.kmi.open.ac.uk/open_dataset. The dataset itself is a relational database containing information on courses, students, and their interactions with Virtual Learning Environment for seven selected courses in the United Kingdom. The dataset was already cleaned without null values in any of the columns. There were also no glaringly obvious outliers when starting to work with the data.

Data Problem

With this dataset of student demographic information and their final result in the course, it seems linear to use this information to predict how students will perform in the course in the future. This type of information seems useful for any online learning company to know in order to increase enrollment rates and student retention.

Some questions this dataset brought up:

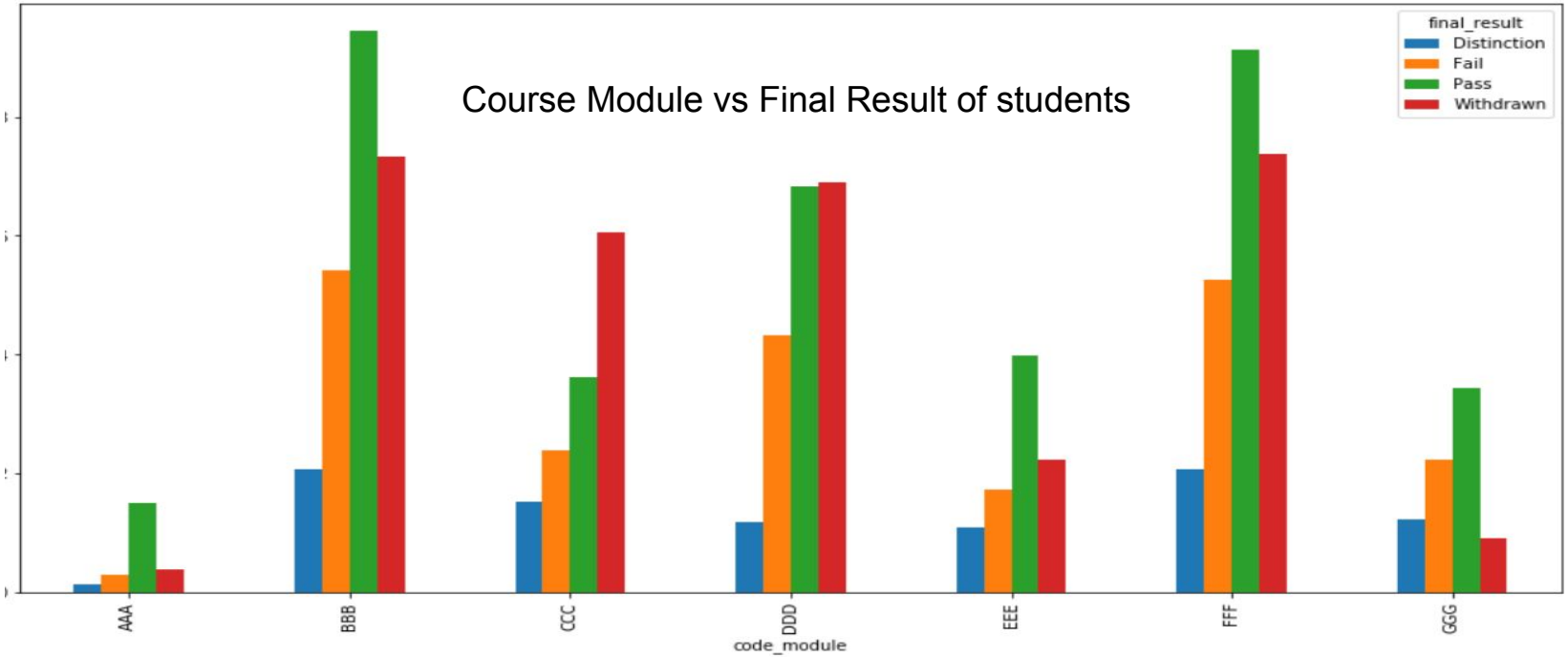
- Can we predict student performance in the course based on demographic information?
- Can we predict student performance in the course based on interaction with the material?
- Can we intervene for students at risk of withdrawing from the course?

Data Wrangling

- Any extraneous data was removed
  - Columns containing only redundant data
  - Columns that held only descriptive information on the course like identifier ID’s
  - Features such as student ID and course information
- Features for Data Exploration and Machine Learning were identified and prepared
  - Mapping ordinal values to integers
  - Encoding nominal values as dummy variables
  - Preparing the target column
  - Summing the number of clicks students made

Data Exploration

There are quite a few variables that are of interest and can have an effect on the student's final result in the course. Various independent variables were explored by examining correlations between independent variables and plotting bar graphs of those variables versus the final result of the student in the course.



Correlation between Independent Variables

	id_student	highest_education	imd_band	age_band
id_student	1.000000	0.042191	0.024396	0.198493
highest_education	0.042191	1.000000	0.125717	0.110026
imd_band	0.024396	0.125717	1.000000	0.071579
age_band	0.198493	0.110026	0.071579	1.000000
num_of_prev_attempts	0.012470	-0.041709	-0.040758	0.005033
studied_credits	-0.006083	0.010024	-0.037872	-0.075143
Distinction	0.028164	0.127624	0.086055	0.064893
Fail	-0.032051	-0.098357	-0.082725	-0.053291
Pass	0.003417	0.062797	0.070033	0.029214
Withdrawn	0.007341	-0.064971	-0.059340	-0.026833
sum_click	0.037327	0.082098	0.077465	0.140429

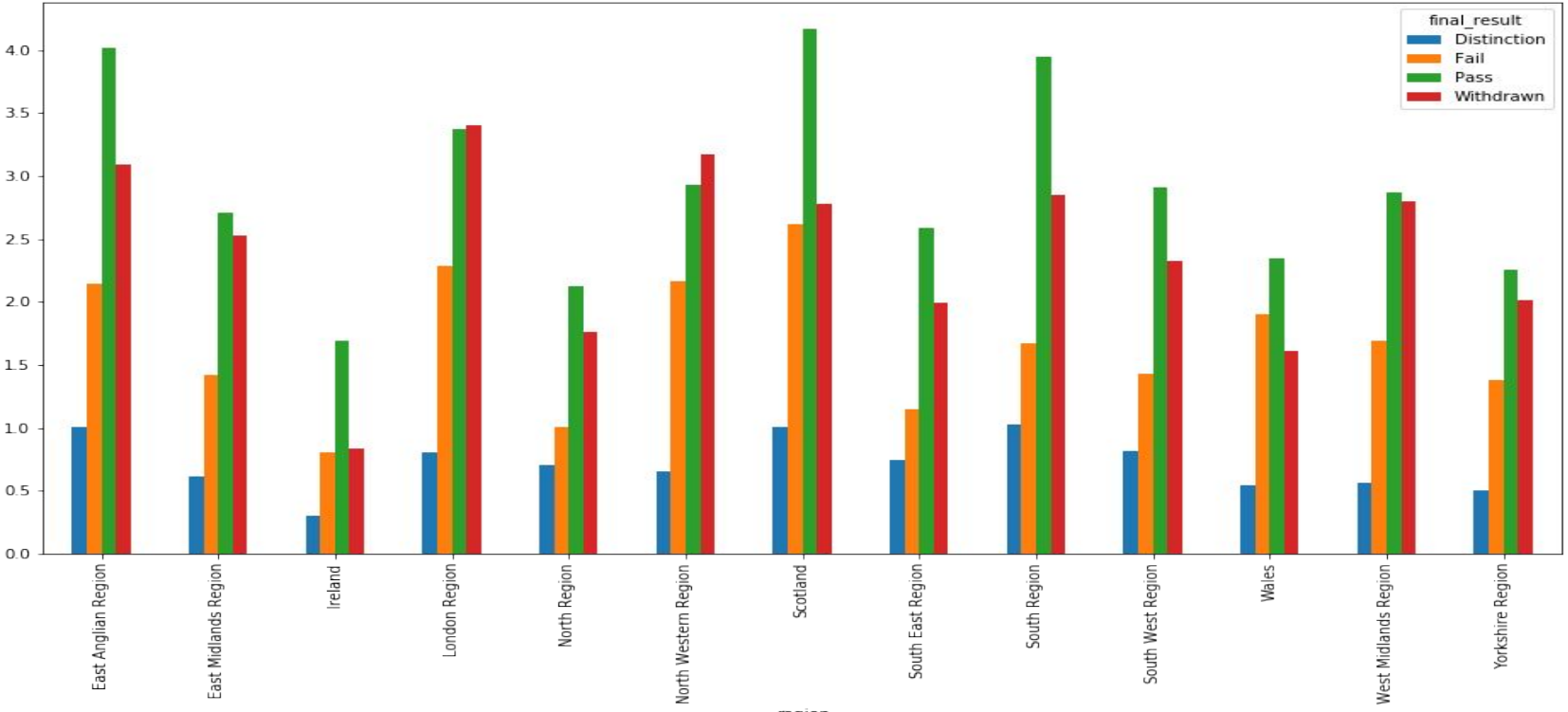
	num_of_prev_attempts	studied_credits	Distinction
id_student	0.012470	-0.006083	0.028164
highest_education	-0.041709	0.010024	0.127624
imd_band	-0.040758	-0.037872	0.086055
age_band	0.005033	-0.075143	0.064893
num_of_prev_attempts	1.000000	0.181726	-0.068826
studied_credits	0.181726	1.000000	-0.057721
Distinction	-0.068826	-0.057721	1.000000
Fail	0.094794	-0.029222	-0.181751
Pass	-0.073689	-0.092765	-0.283667
Withdrawn	0.039994	0.172529	-0.197805
sum_click	-0.068865	-0.006473	0.257051

	Fail	Pass	Withdrawn	sum_click
id_student	-0.032051	0.003417	0.007341	0.037327
highest_education	-0.098357	0.062797	-0.064971	0.082098
imd_band	-0.082725	0.070033	-0.059340	0.077465
age_band	-0.053291	0.029214	-0.026833	0.140429
num_of_prev_attempts	0.094794	-0.073689	0.039994	-0.068865
studied_credits	-0.029222	-0.092765	0.172529	-0.006473
Distinction	-0.181751	-0.283667	-0.197805	0.257051
Fail	1.000000	-0.455504	-0.317630	-0.209286
Pass	-0.455504	1.000000	-0.495738	0.280061
Withdrawn	-0.317630	-0.495738	1.000000	-0.299305
sum_click	-0.209286	0.280061	-0.299305	1.000000

- The feature variables appear to be independent
- There appears to be a correlation between the number of clicks a student makes and their final result in the course
- This leads us to a data question:
  - Can we predict withdrawal of a student based on these independent variables?

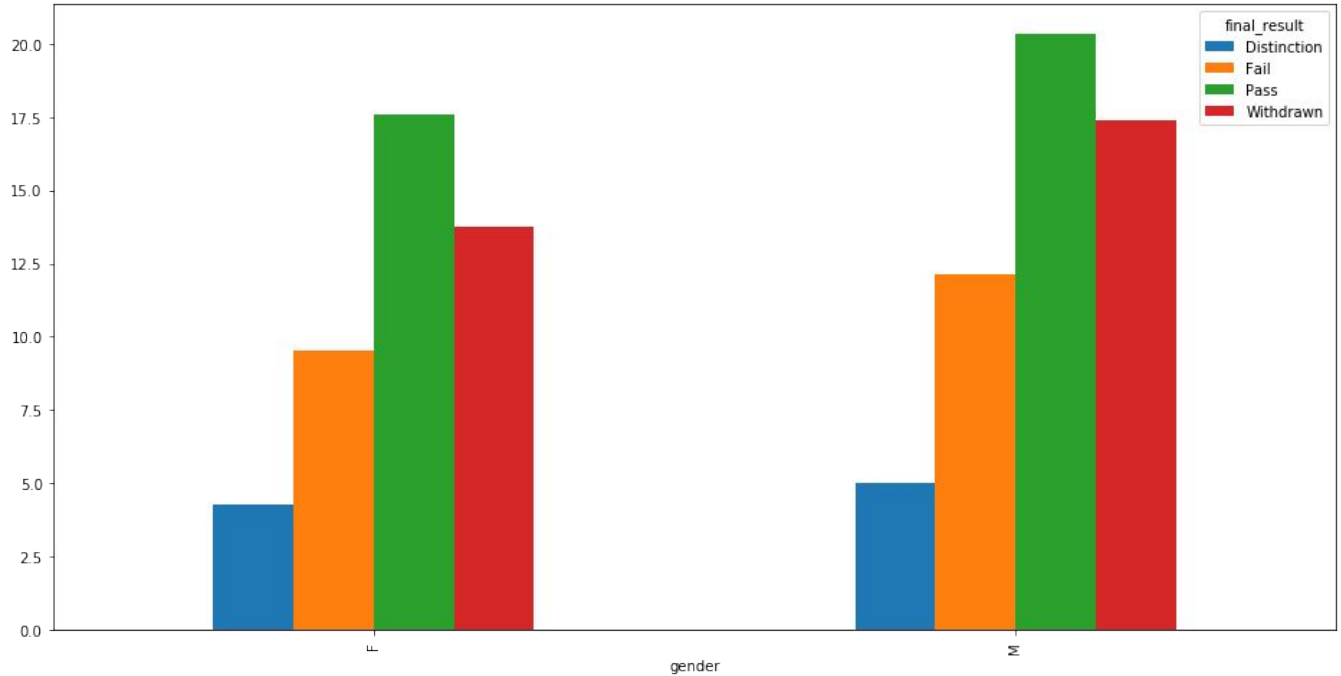
Data Exploration - Findings

Students are distributed relatively evenly over different geographic regions.



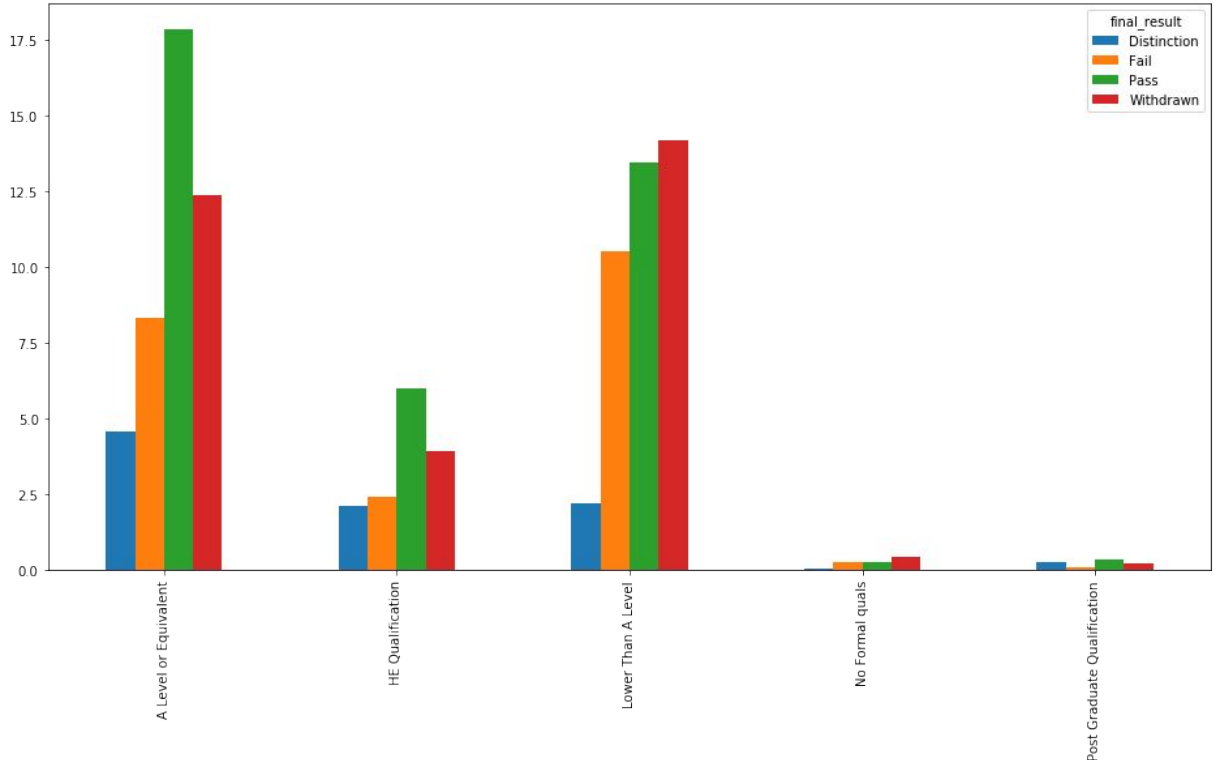
Data Exploration - Findings

- Men and women appear to have similar pass and failure rates



Data Exploration- Findings

Most of the people taking the course have A level (secondary school level) education or lower.



Data Exploration - Findings

- 31.1% of the students withdrew from the courses and 21.6% failed out of the courses.
- A total of 52.7% failed, implying a 47.3% successful rate, i.e., slightly less than half of the students registering for the courses made it through to the end.
- to increase the successful pass rate, proactive, targeted, and timely intervention will be critically important.

Data Analysis - Machine Learning

To prepare for machine learning, a label encoder was applied to the dataset to turn string variables into scalars. Then a one hot encoding was done to create binary variables out of the scalar variables for machine learning modeling.

n [197]:
```
#Create Dummy Variable to isolate students that Withdrew from the course

combined['final']= np.where(combined['final_result']=='Withdrawn', 1, 0)
```

n [198]:
```
#Label Encoder to turn strings into scalars

le = preprocessing.LabelEncoder()
combined = combined.apply(le.fit_transform)
```

n [199]:
```
#One hot Encoding

enc = OneHotEncoder(sparse='True')
columns_to_encode = ['region','highest_education','imd_band','age_band','num_of_prev_attempts',]
```

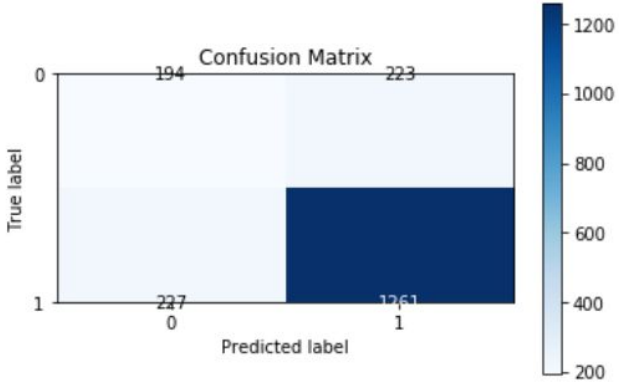
n [200]:
```
def one_hot(df, cols):
    for each in cols:
        dummies = pd.get_dummies(df[each], prefix=each, drop_first=False)
        df = pd.concat([df, dummies], axis=1)
    return df
```

n [201]:
```
#one hot encoding

combined=one_hot(combined,columns_to_encode)
```

Machine Learning

As the goal of the data analysis is to classify students at risk of withdrawal from the course, a supervised machine learning classification method seems appropriate. A Decision Tree model and a Random Forest Model were built on the dataset. After cross validating on four folds the models we built on the training set to prevent overfitting, it was discovered that the random forest model had the highest precision in predicting withdrawal rates of students in the course.

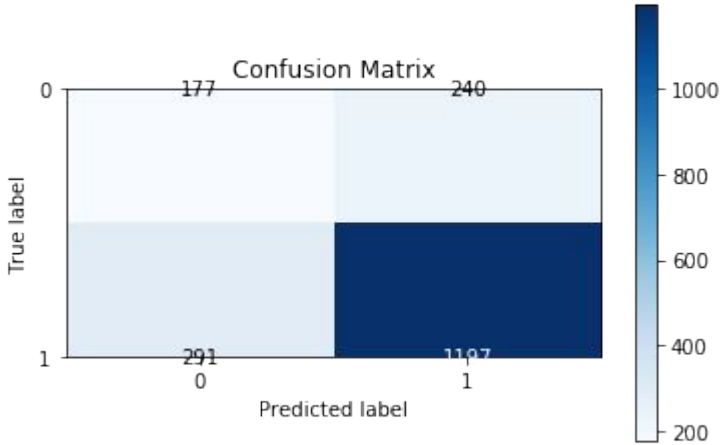


- Confusion matrix for random forest showing that the model predicted true positives at a high rate of precision

```
print("Precision:{0:.3f}".format(metrics.precision_score(yTrain, y_pred)),"\n")
```

```
Precision:0.850
```

Machine Learning - Confusion Matrix



From this confusion matrix for the cross validated decision tree model, we can see that it performs slightly worse than the random forest model. Precision is 0.833 vs 0.850 for the random forest model

Machine Learning

After choosing the random forest model built on the training set. We used the test set to check to see how our model was performing and compared the dataset containing just demographic information with the full dataset containing demographic information and the number of clicks students made.

```
# Random Forest for Table with sum_click
X = combined.loc[:, combined.columns != 'final']
y = combined['final']
xTrain, xTest, yTrain, yTest = train_test_split(X, y,train_size = 0.75)

rf = RandomForestClassifier(n_estimators=10,random_state=33)
rf = rf.fit(xTrain, yTrain)
train_pred = rf.predict(xTrain)
test_pred = rf.predict(xTest)
print("Precision:{0:.3f}".format(metrics.precision_score(yTest, test_pred)),"\n")
```

```
Precision:0.852
```

```
# Random Forest for Table without sum_click
X = combinednosum.loc[:, combinednosum.columns != 'final']
y = combinednosum['final']
xTrain, xTest, yTrain, yTest = train_test_split(X, y,train_size = 0.75)
```

Results

- The model that was built using demographic information and number of clicks students made had the highest positive predictive value of them all
- This result falls in line with our reasoning as those students who clicked more and interacted more with the course material were less likely to withdraw from the course.
- There was a 0.7% improvement from 84.5% to 85.2% in the performance of the model when including the number of clicks students made.

Conclusion

With the machine learning model built in this project, students at risk of withdrawing from the course were identified and intervention could then be conducted to help these students graduate from the course.

This information is useful for any online teaching company that wishes to improve their user experience. Engagement of any kind from the student with course material increased that student’s probability of passing the course.

In the future, more experiments could be conducted to find other independent variables that have a positive effect on student’s matriculation rates.