CSDA1050 Advanced Analytics Capstone Course

Improving student's graduation in Education

Sylvain Kamto

Student ID: 11060

1- Introduction/ Background

- Currently less than 65% of the students complete their studies as planned. Part of the students will move to work without graduation or change the branch of studies to another institute, but too many have either delayed in their studies (12.3%) or will completely discontinue (8.5%)

- The delayed and dropout students pose significant direct costs to cities and schools due to reduced funding from government. Dropouts especially have challenges in finding a job and this problem is causing serious impacts on society in the long run.

- To alleviate this problem, we are here by initiating a concept project on how to apply analytics to improve graduation in schools. The core of the idea is the following: utilize advanced analytics and machine learning to identify students who have elevated risk to dropout or delay in studies, so that interventions and support actions can be initiated early enough.

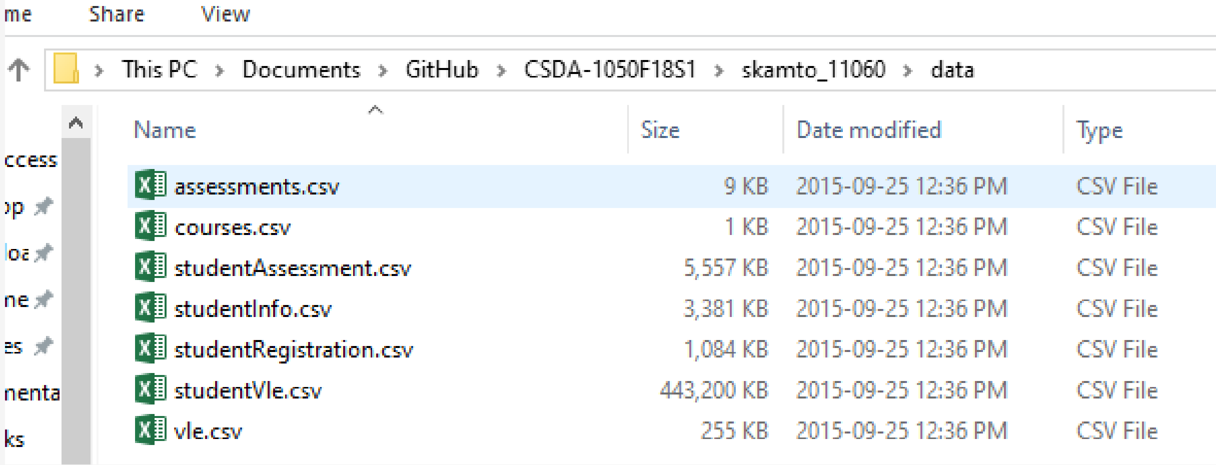
2- Research Question

1. Predicts which students have elevated risk of delayed studies or even dropping out
2. Predict student academic outcomes to better guidance and support

3- Dataset

Data were collected from the anonymised Open University Learning Analytics Dataset (OULAD). It contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules)

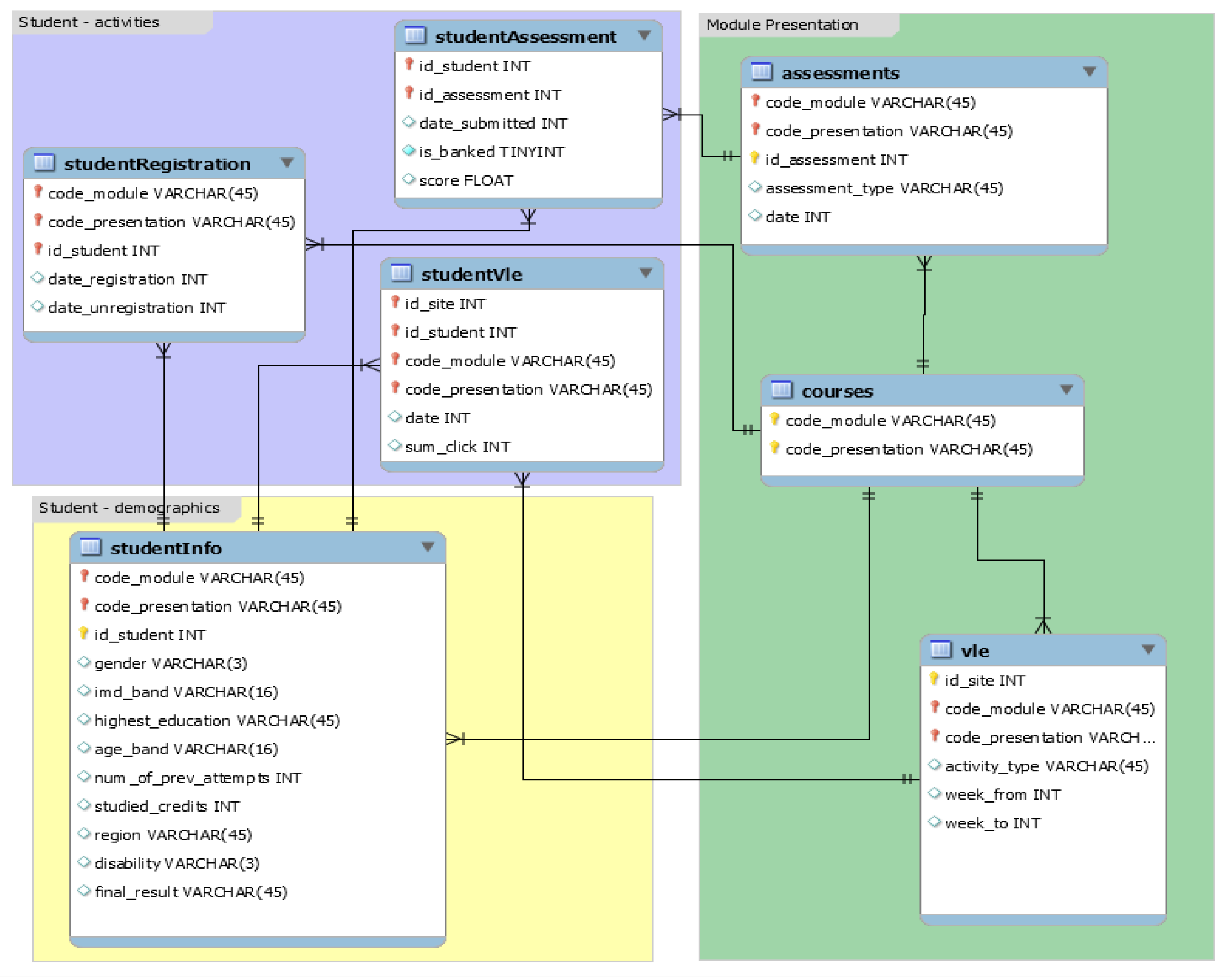
The dataset consists of tables connected using unique identifiers. All tables are stored in the csv format.


me
Share
View
This PC › Documents › GitHub › CSDA-1050F18S1 › skamto_11060 › data
ccess
op
loa
ne
es
nenta
ks

Name	Size	Date modified	Type
assessments.csv	9 KB	2015-09-25 12:36 PM	CSV File
courses.csv	1 KB	2015-09-25 12:36 PM	CSV File
studentAssessment.csv	5,557 KB	2015-09-25 12:36 PM	CSV File
studentInfo.csv	3,381 KB	2015-09-25 12:36 PM	CSV File
studentRegistration.csv	1,084 KB	2015-09-25 12:36 PM	CSV File
studentVle.csv	443,200 KB	2015-09-25 12:36 PM	CSV File
vle.csv	255 KB	2015-09-25 12:36 PM	CSV File

4- Dataset Description

This dataset offers two of the elements in the framework: behavior and performance. It contains information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries).



Student - activities

studentAssessment
- id_student INT
- id_assessment INT
- date_submitted INT
- is_banked TINYINT
- score FLOAT

studentRegistration
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_student INT
- date_registration INT
- date_unregistration INT

studentVle
- id_site INT
- id_student INT
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- date INT
- sum_click INT

Student - demographics

studentInfo
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_student INT
- gender VARCHAR(3)
- imd_band VARCHAR(16)
- highest_education VARCHAR(45)
- age_band VARCHAR(16)
- num_of_prev_attempts INT
- studied_credits INT
- region VARCHAR(45)
- disability VARCHAR(3)
- final_result VARCHAR(45)

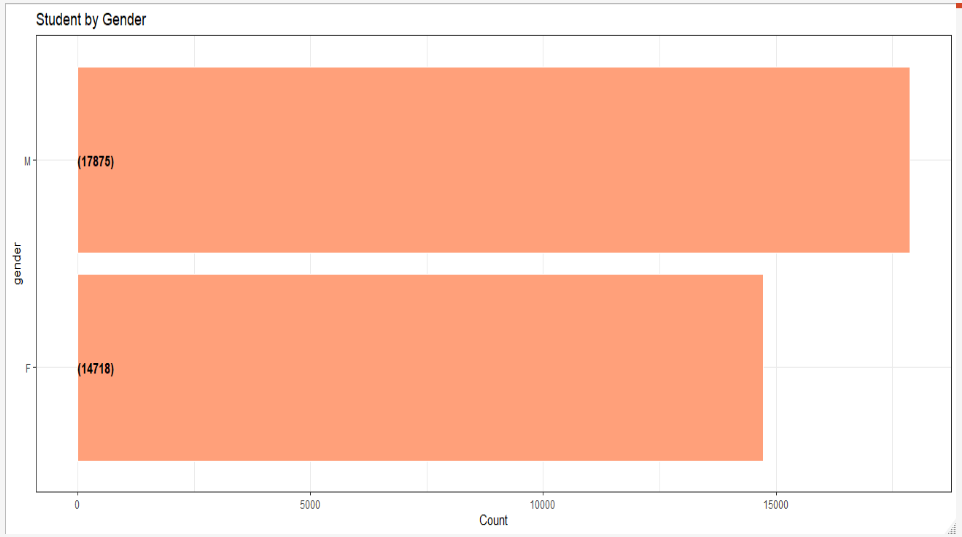
Module Presentation

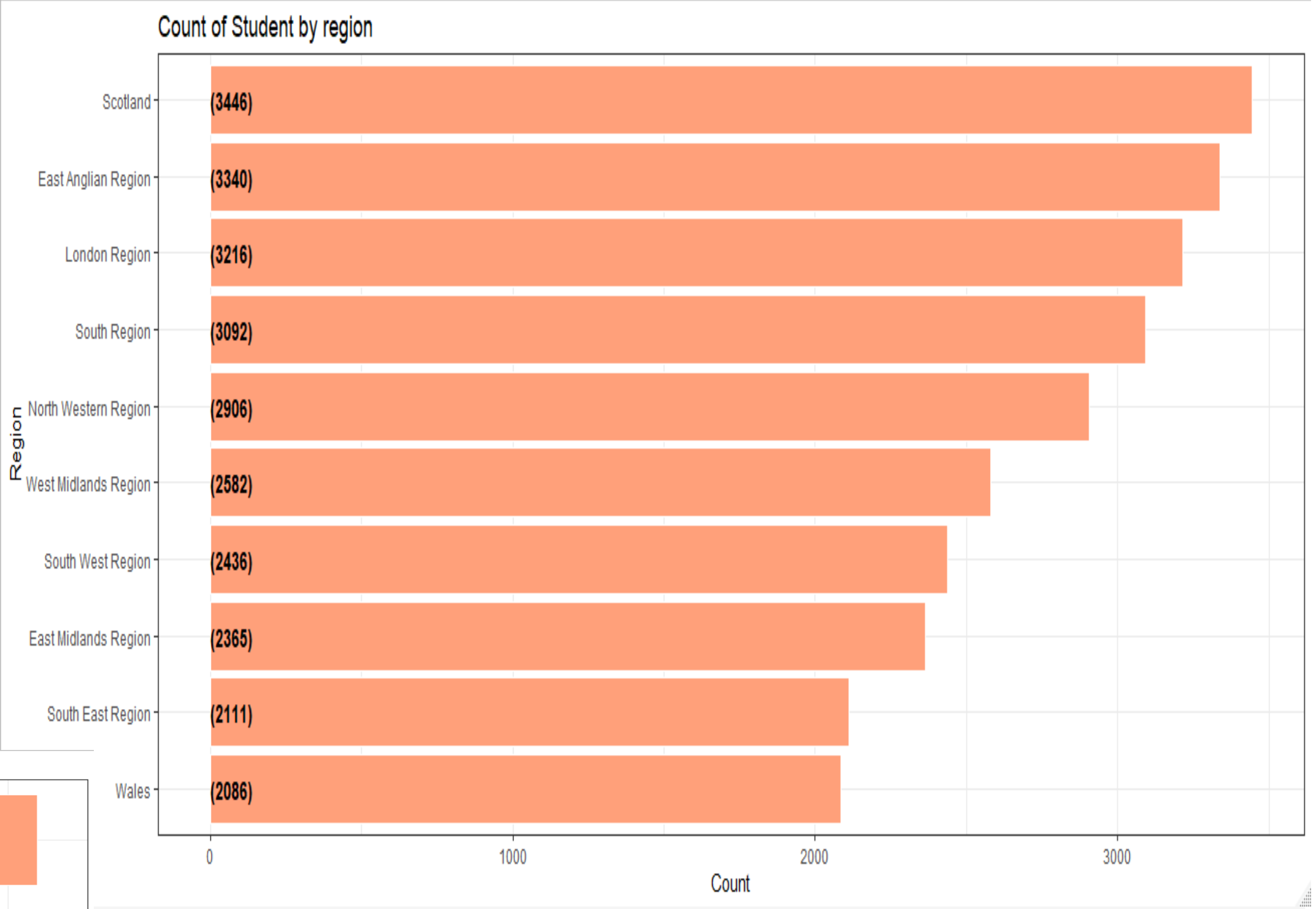
assessments
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_assessment INT
- assessment_type VARCHAR(45)
- date INT

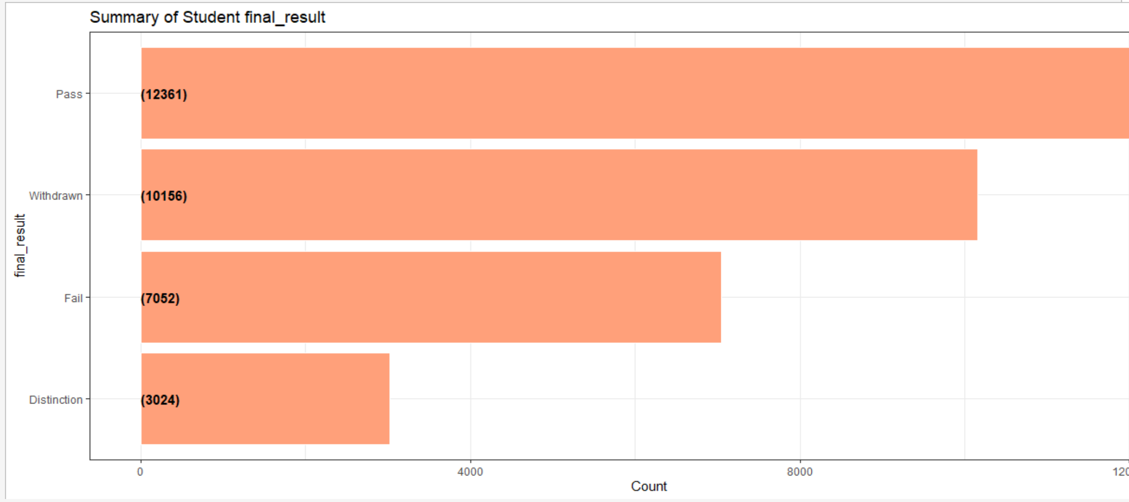
courses
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)

vle
- id_site INT
- code_module VARCHAR(45)
- code_presentation VARCH...
- activity_type VARCHAR(45)
- week_from INT
- week_to INT

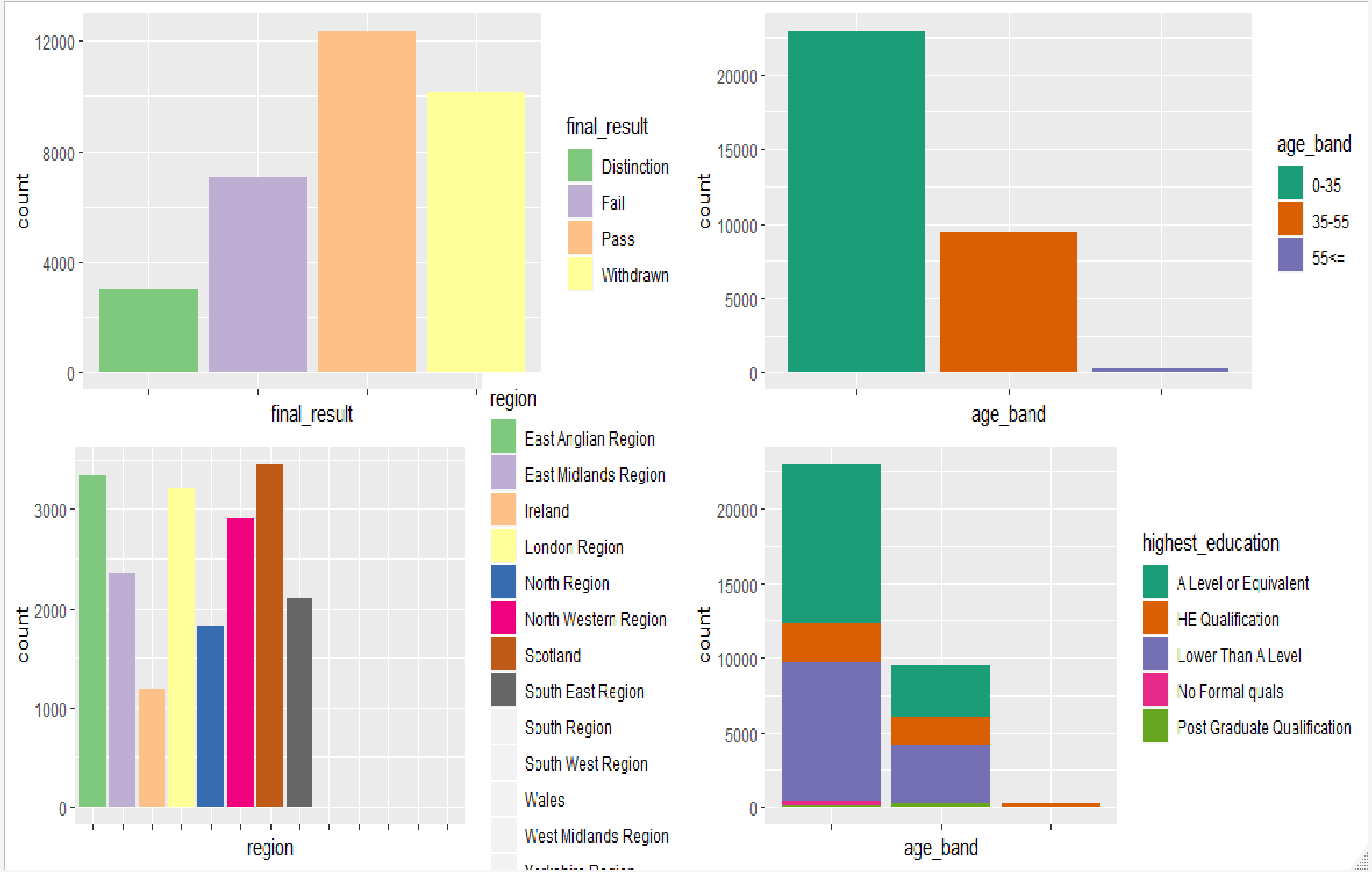
Data exploration







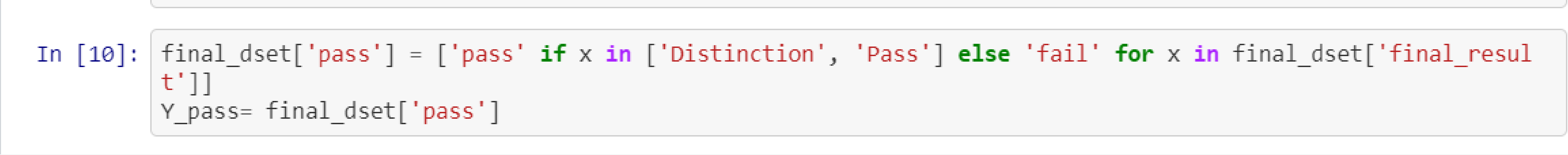
Data exploration



Analytics

Target Variable :

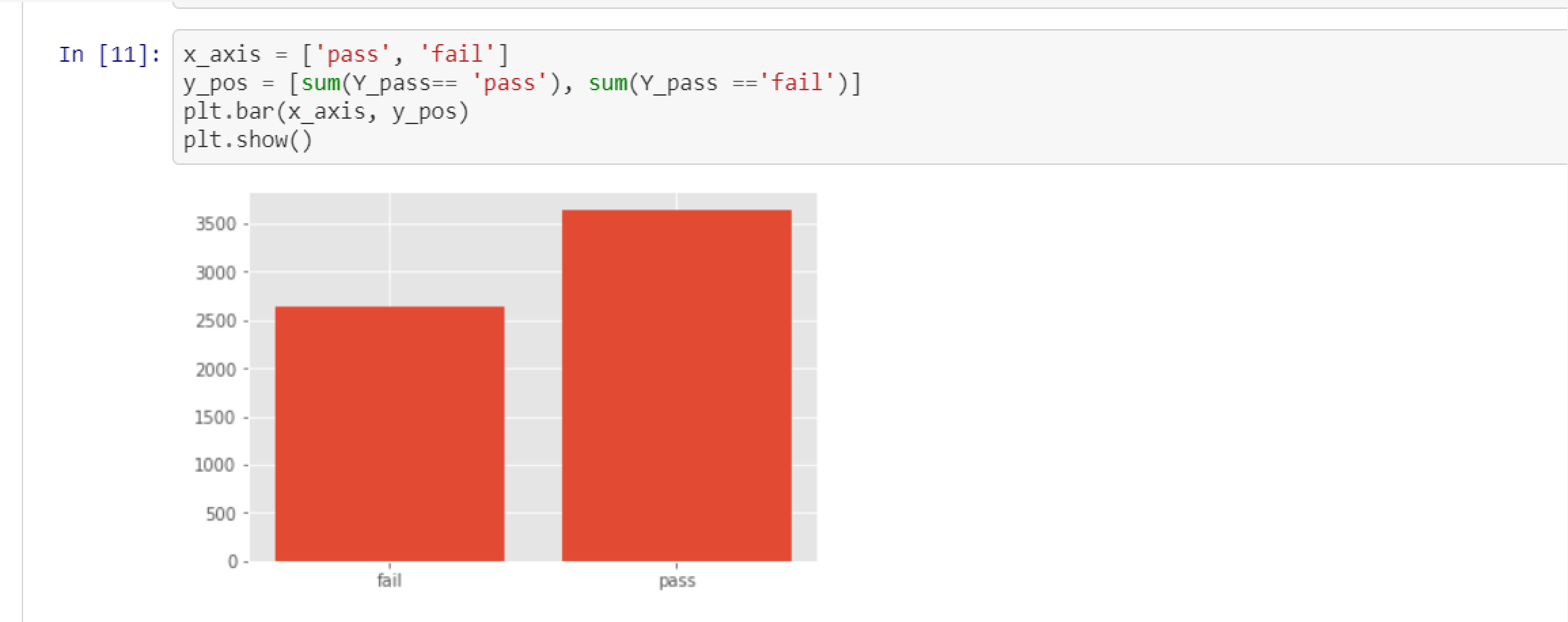
- Fail or Pass



In [10]:

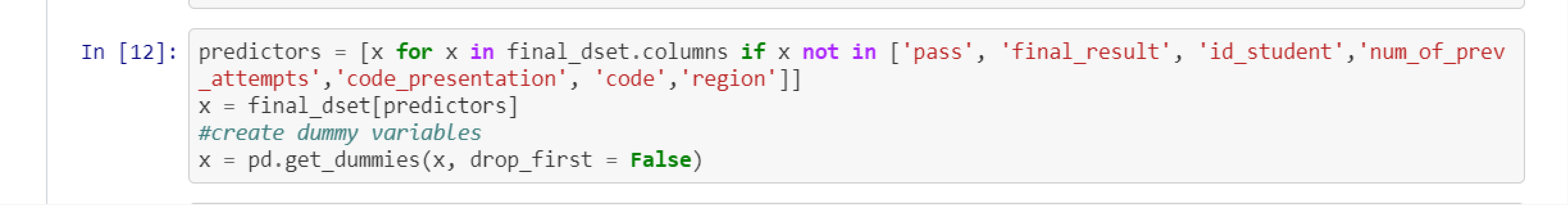
```
final_dset['pass'] = ['pass' if x in ['Distinction', 'Pass'] else 'fail' for x in final_dset['final_result']]
Y_pass= final_dset['pass']
```

Bar Chart of distribution of students who failed and passed



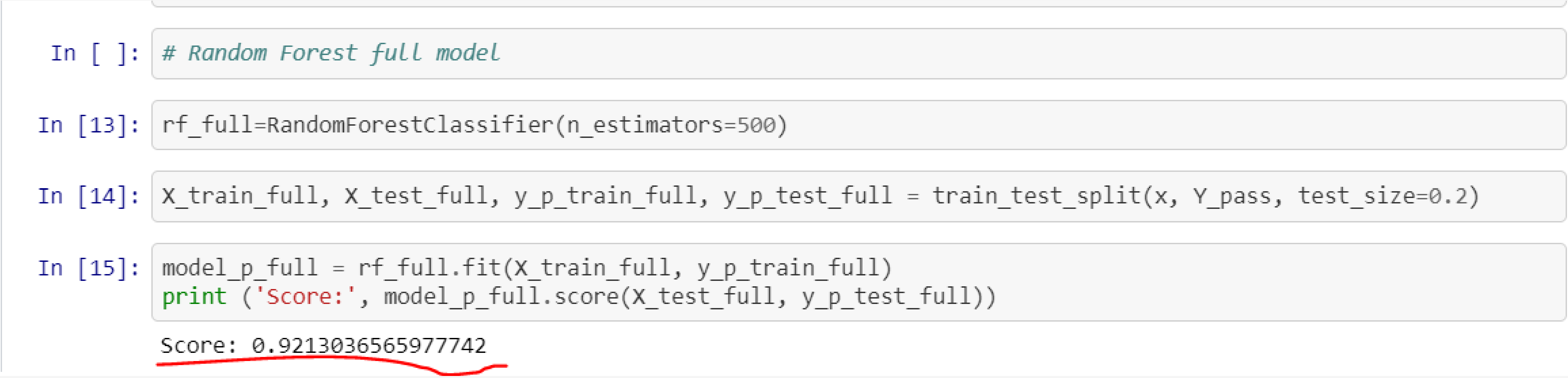
```
In [11]: x_axis = ['pass', 'fail']
         y_pos = [sum(Y_pass== 'pass'), sum(Y_pass =='fail')]
         plt.bar(x_axis, y_pos)
         plt.show()
```

Predictor variable



```
In [12]: predictors = [x for x in final_dset.columns if x not in ['pass', 'final_result', 'id_student','num_of_prev_attempts','code_presentation', 'code','region']]
         x = final_dset[predictors]
         #create dummy variables
         x = pd.get_dummies(x, drop_first = False)
```

Random Forest full Model



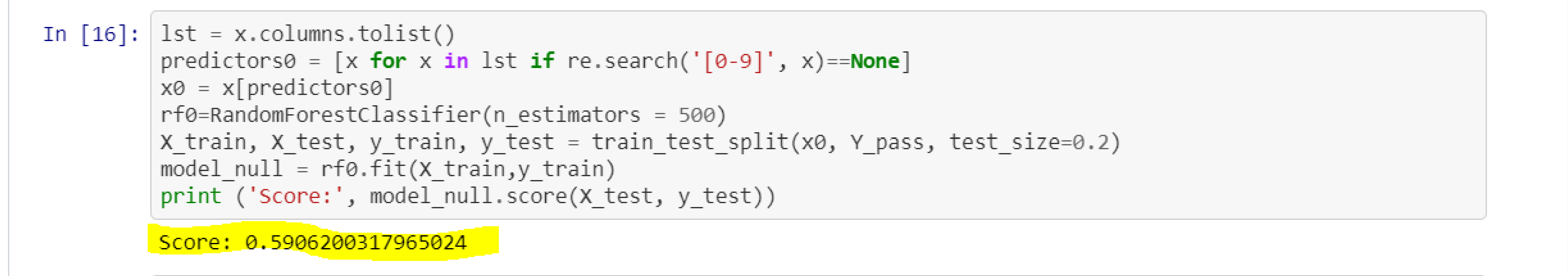
```
In [ ]: # Random Forest full model

In [13]: rf_full=RandomForestClassifier(n_estimators=500)

In [14]: X_train_full, X_test_full, y_p_train_full, y_p_test_full = train_test_split(x, Y_pass, test_size=0.2)

In [15]: model_p_full = rf_full.fit(X_train_full, y_p_train_full)
         print ('Score:', model_p_full.score(X_test_full, y_p_test_full))

         Score: 0.9213036565977742
```

The full model with all of the students assessment scores, weekly virtual interaction scores, etc. Is fairly robust at predicting student failure. But this prediction is made with all data available. Or over 30 weeks of data.

Can student success of failure be predicted earlier?

Model with only demographic data

this is data of only student demographic data. So the model does not use any course information



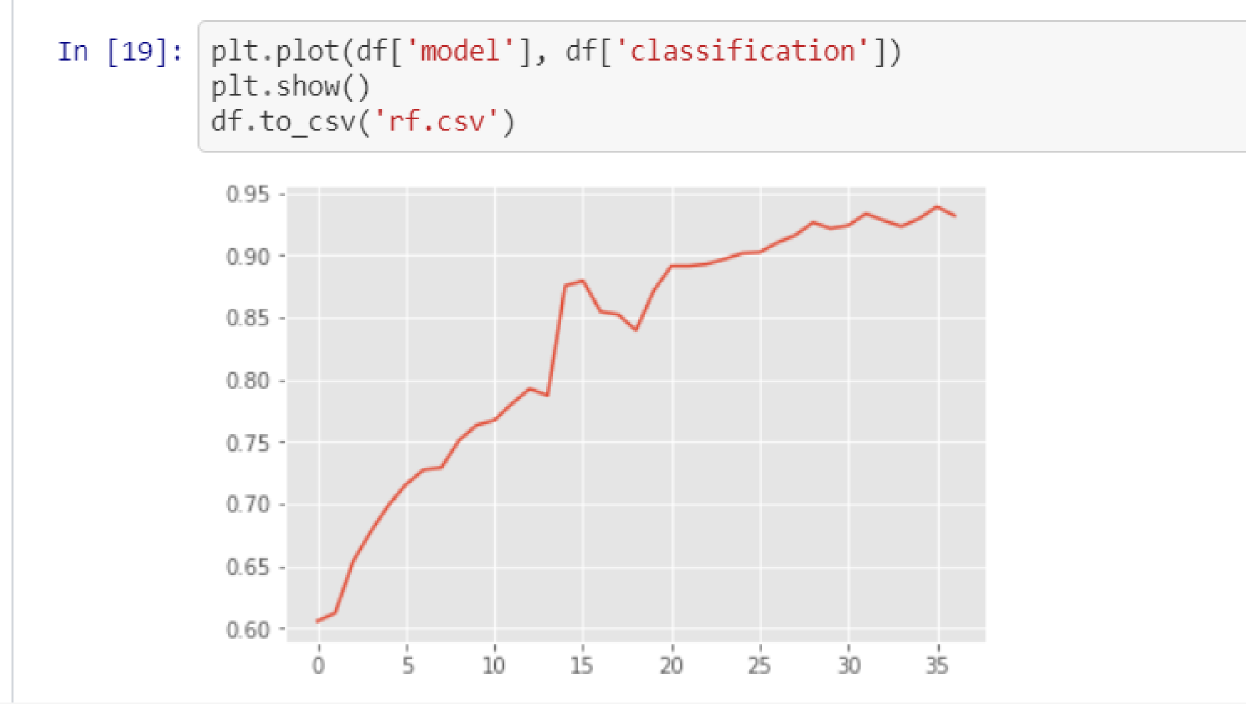
In [16]:

```
lst = x.columns.tolist()
predictors0 = [x for x in lst if re.search('[0-9]', x)==None]
x0 = x[predictors0]
rf0=RandomForestClassifier(n_estimators = 500)
X_train, X_test, y_train, y_test = train_test_split(x0, Y_pass, test_size=0.2)
model_null = rf0.fit(X_train,y_train)
print ('Score:', model_null.score(X_test, y_test))
```

```
Score: 0.5906200317965024
```

now with only demographic data, the model can predict success of failure with about 60% accuracy

Steady increase in model prediction accuracy with every additional week of data. However, we see jumps at certain weeks. This coincides with weeks where student had tests

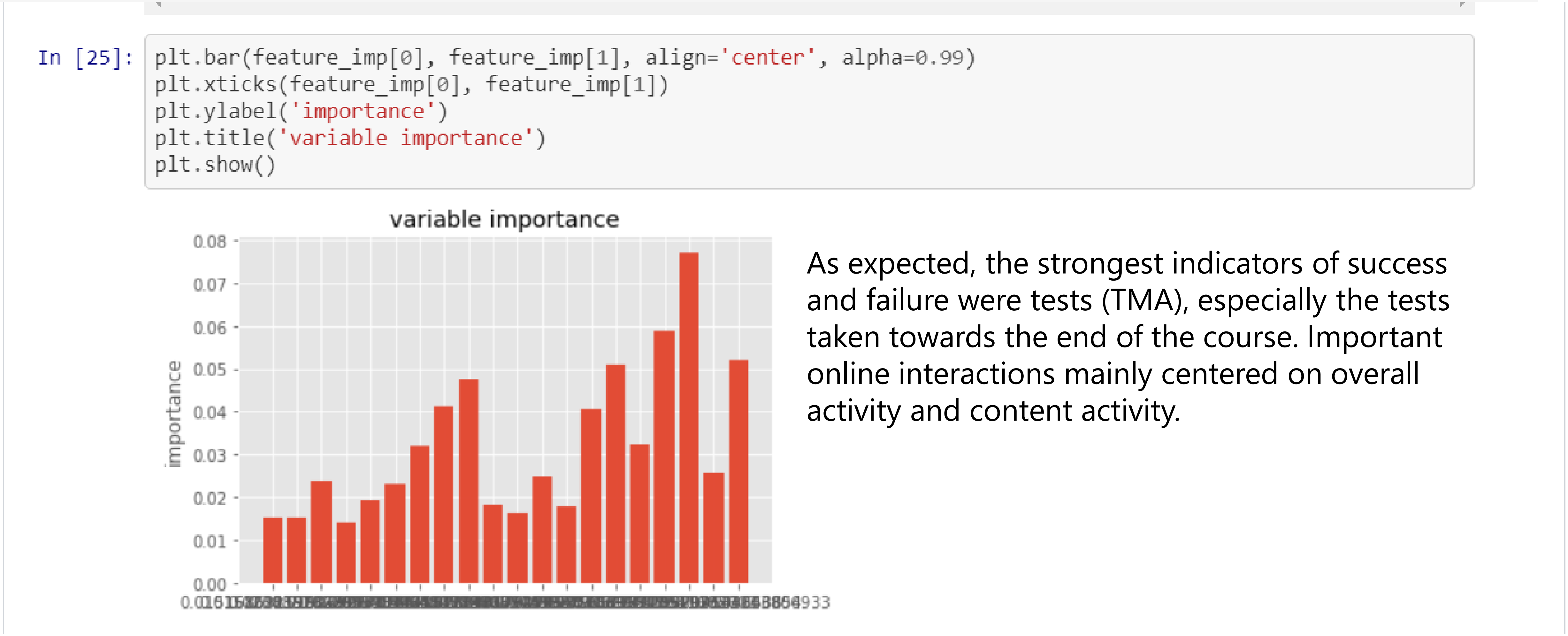


```
plt.plot(df['model'], df['classification'])
plt.show()
df.to_csv('rf.csv')
```

by week 8, the model could predict with 75% accuracy student failure. Over 85% accuracy by week 13. And after week 24, there seems to be marginal gains, suggesting that additional student information does not add to model predictive power

Feature importance

The model can predict with confidence student success and failure, but what features does the model identify as important for its prediction?



```
In [25]: plt.bar(feature_imp[0], feature_imp[1], align='center', alpha=0.99)
         plt.xticks(feature_imp[0], feature_imp[1])
         plt.ylabel('importance')
         plt.title('variable importance')
         plt.show()
```

As expected, the strongest indicators of success and failure were tests (TMA), especially the tests taken towards the end of the course. Important online interactions mainly centered on overall activity and content activity.

Conclusion

- As online learning continues to grow as a platform to educate students, it is important to consider how we can use the data associated with these programs to best identify students who are at risk for failure.

- I use random forests to see how early student failure can be predicted. The models showed that with demographic information alone, student failure can be predicted with 60% accuracy. As weekly information on student online interactions and assessments are added, predictions for students at risk for failure greatly increase. By week 8, the model improves to 75% accurcacy, and has over 85% accuracy by week 13.

- As educators continue to seek ways to identify students at risk for failure, this model provides a robust way to do so. However, it is not enough to find students at risk for failure. Intervention needs to be implemented to best understand how to improve the outcomes of these students.