Data Exploration from OKCupid

- Given the personal profiles of users of OKCupid, we attempt to learn and predict interesting features from these data

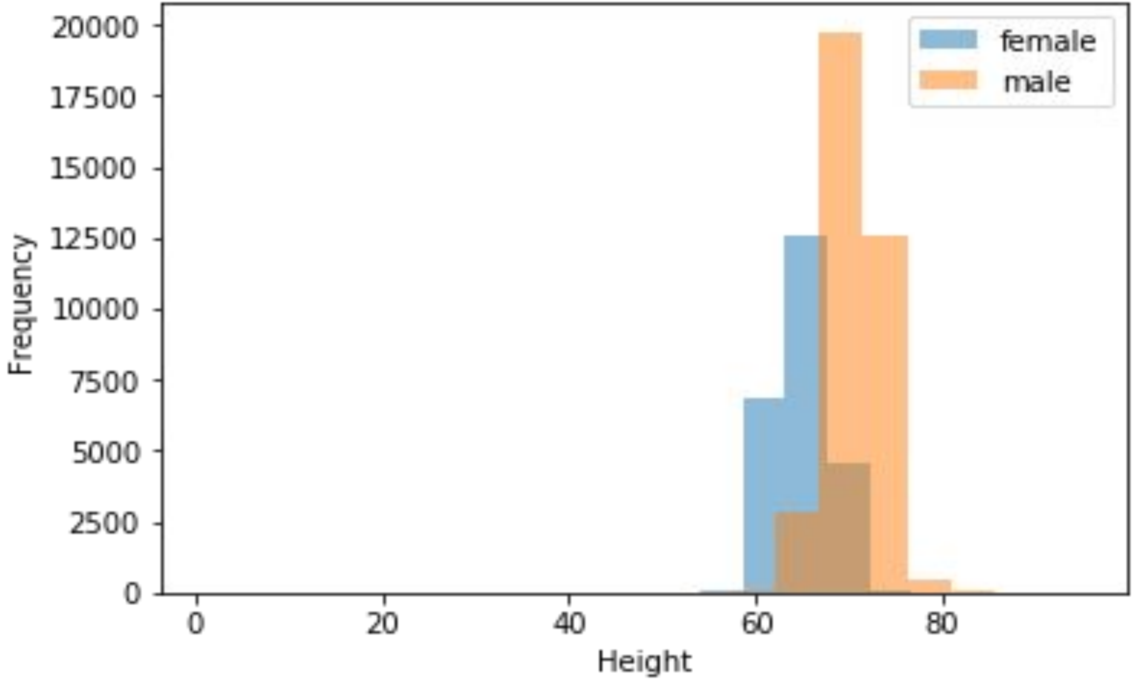
Machine Learning Fundamentals
Ming Ming Tan

Table of Contents

- Exploration of the Dataset
- Questions to Answer
- Augmenting the Dataset
- Classification Approaches
- Regression Approaches
- Conclusions/Next steps

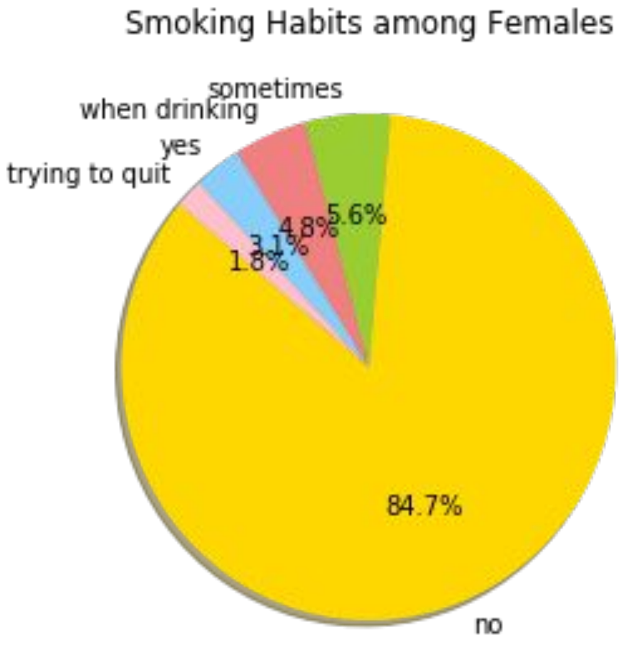
Exploration of Dataset on Gender using Heights and Smoking Habits

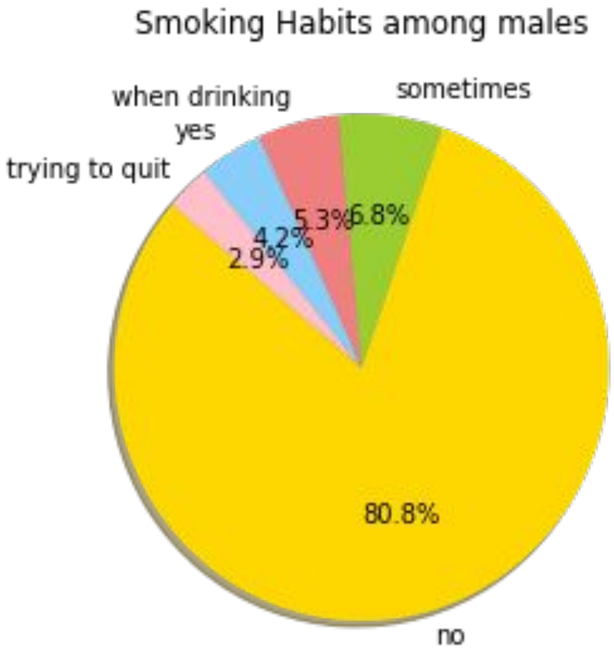
Frequency of Heights between Different Genders



From the graph, it appears that males generally have heights above 65 while majority of the females have heights less than 65.

Percentage of Smoking Habits between Different Genders





From the pie chart, it appears that both males and females have similar distribution of smoking habits.

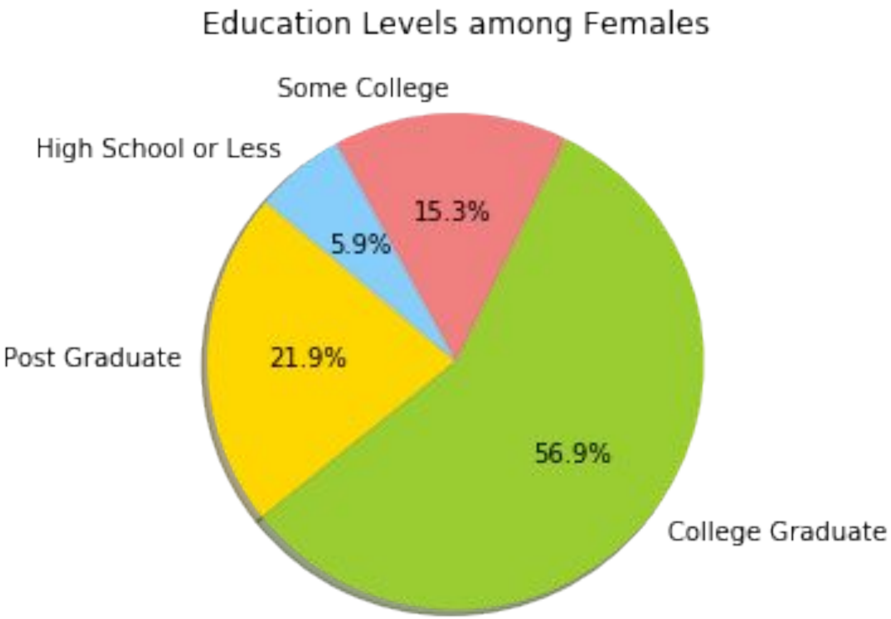
Exploration of Dataset on Education

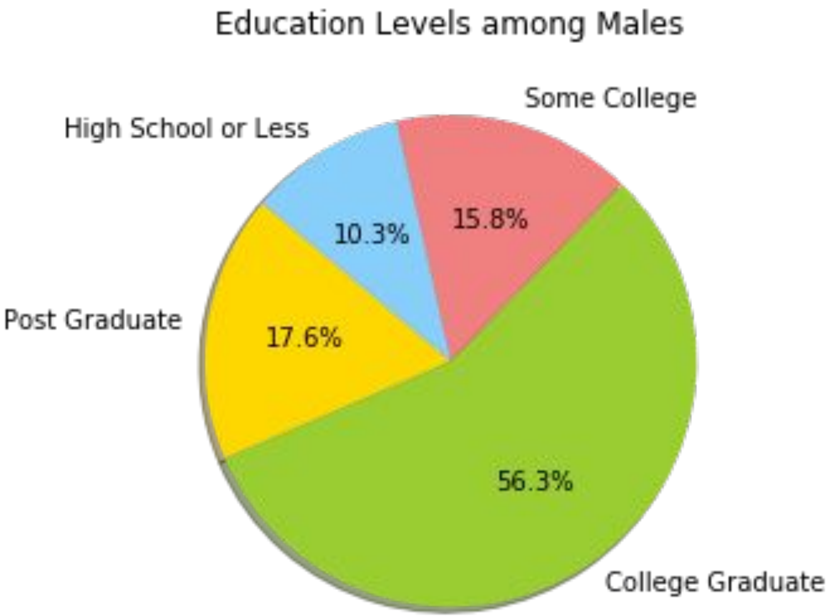
Classification of Education Levels

- we classify users’ response to the question on education into four broad categories

Post graduate	College graduate	Some college	High school of less
'graduated from masters program' 'graduated from ph.d program'	'graduated from college/university', 'working on masters program', ‘graduated from two-year college’, 'working on ph.d program’, 'dropped out of masters program', masters program', 'dropped out of ph.d program', 'ph.d program' 'graduated from law school' 'graduated from med school'	'working on college/university' 'working on two-year college' 'college/university' 'working on law school' 'two-year college' 'working on med school' 'law school' 'med school'	'graduated from high school', 'dropped out of college/university', 'graduated from space camp', 'dropped out of space camp', 'working on space camp', 'dropped out of two-year college', ‘'dropped out of high school', 'high school', 'working on high school', 'space camp', 'dropped out of law school', 'dropped out of med school'

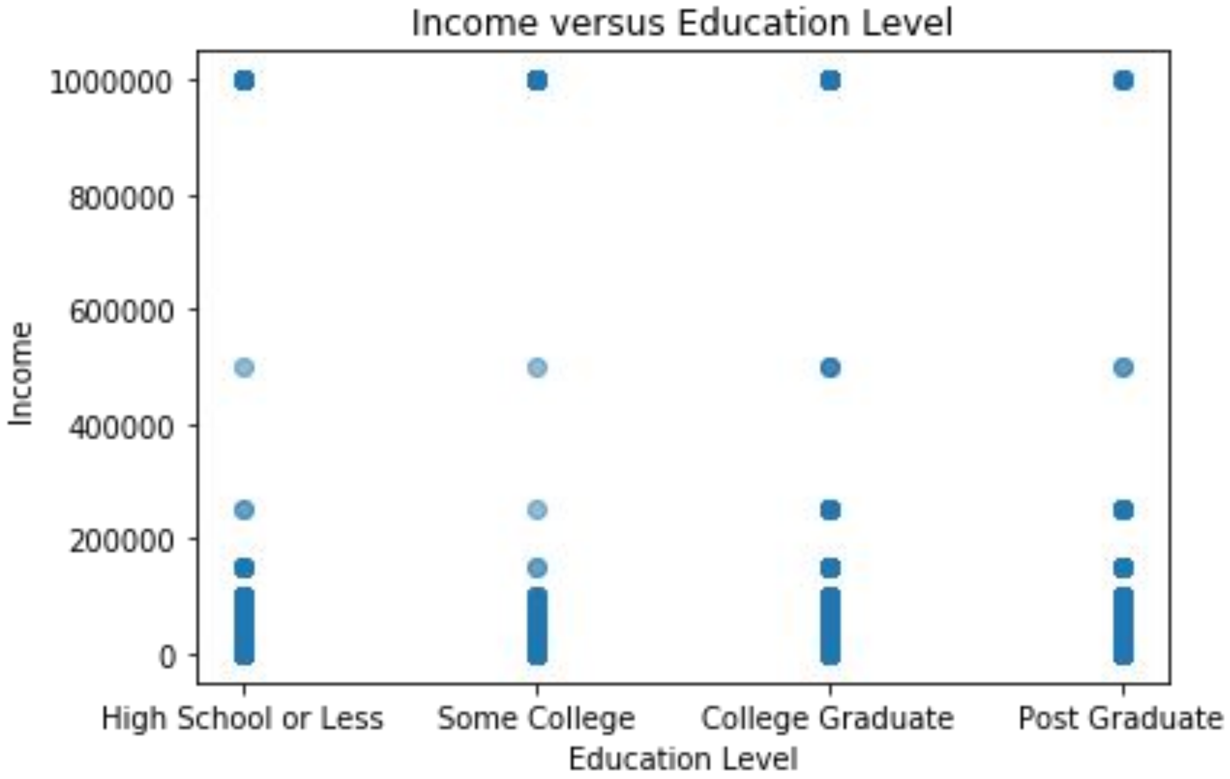
Statistics on Education Levels





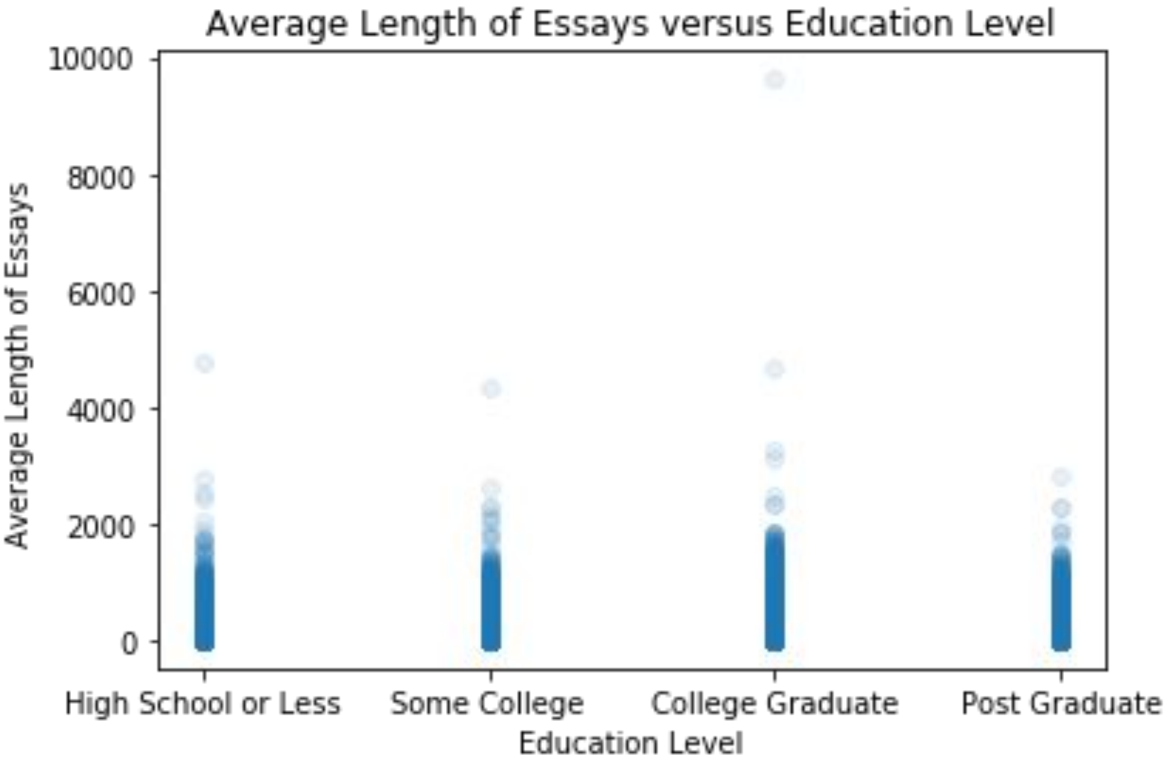
From the pie chart, it appears that both males and females have similar distribution of education levels.

Income versus Education Level



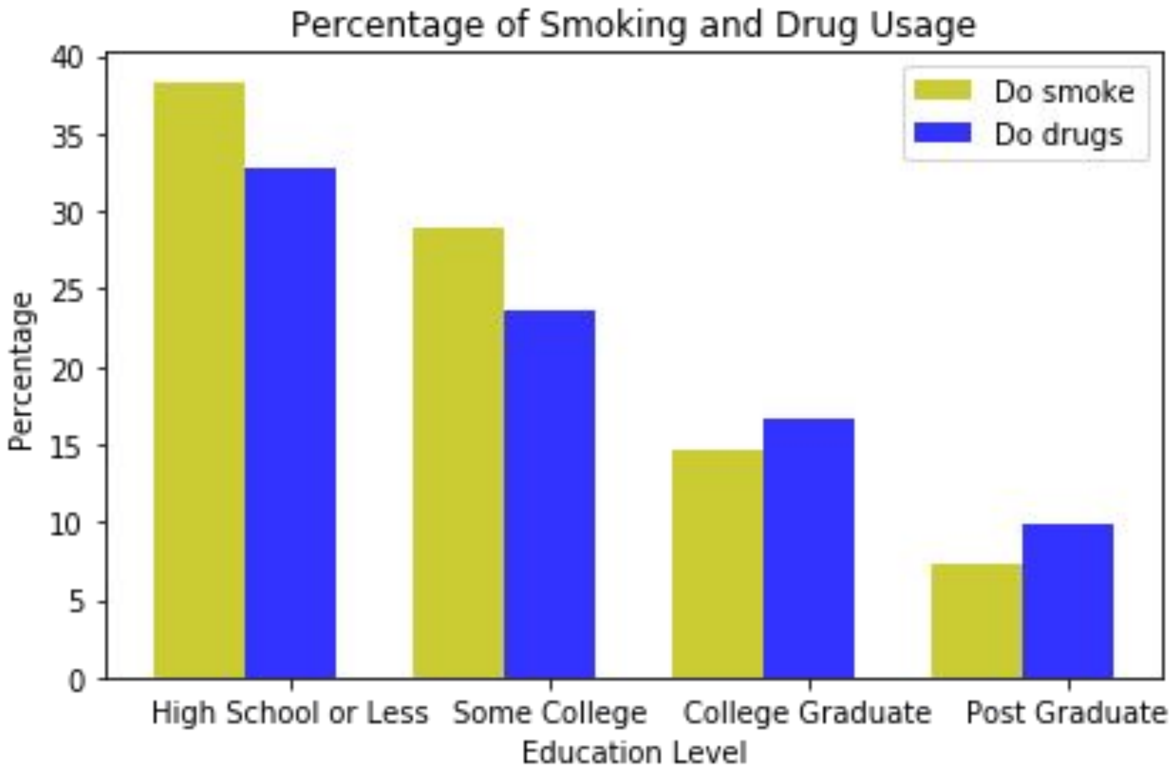
From the graph, it does not appear to have strong linear correlation between education level and income.

Average Length of Essays versus Education Level



From the graph, it does not appear to have strong linear correlation between average length of essays and education level.

Percentage of Smoking and Drug Usage between different Education Level



From the graph, it appears that the higher the education level, the lesser the percentage of users who smoke or do drugs.

Questions to Answer

1. **Can we predict a person's gender using height?**

Slide 4 suggested that classification on gender using height is feasible.

2. **Can we predict a person's education level using features such as income, drugs and smoking usage or average length of essays?**

Slide 8, Slide 9, Slide 10 suggested that each of these features alone may not be sufficient to learn a good prediction. Hence, we will attempt to use combinations of these features to learn the classification of educations levels.

3. **Can we predict a person’s income using combinations of features such as education level, gender etc.**

Slide 8 suggested that the correlation between income and education level is not strong. We will attempt to use combination of different features for the multiple linear regression model.

Augmenting the Dataset

Tables of mapping data to integers

sex	gender_code
‘m’	0
‘f’	1

drinking	drinks_code
‘not at all’	0
‘rarely’	1
‘socially’	2
‘often’	3
‘very often’	4
‘desperately’	5

New column - Mapping of Education Levels

education	education_level
high school or below	0
some college	1
college graduate	2
post graduate	3

We refer to Slide 7: Classification of Education Levels on the four categories of education levels: high school or less, some college, college graduate and post graduate.

New columns - Mapping of Drug and Smoke Usage

drug	do_drug
‘never’	0
‘sometimes’, ‘often’	1

smoke	do_smoke
‘no’	0
‘sometimes’, ‘when drinking’, ‘yes’, ‘trying to quit’	1

New column - Average Length of Essays

- A new column 'essay_ave_len' is created by
  - replacing any ‘nan’ values in all essays with the empty string
  - the value of essay_ave_len for each user is the total length of all essays by the user divided by 10 (the number of essays questions for each user)

- The following describes the summary of statistics of ‘essay_ave_len’:

count	59946
mean	220.907642
std	202.14557
min, max	0, 9, 9627.7

New column - Mapping of Job Status

job	job_status_code
'unemployed'	0
'student'	1
'retired'	2
'military'	3
other than all the above	4

Classification Approaches

KNeighborsClassifier

- We build models using KNeighborsClassifier to classify education level using some combinations of income, essay_ave_len, do_smoke, do_drug and age.

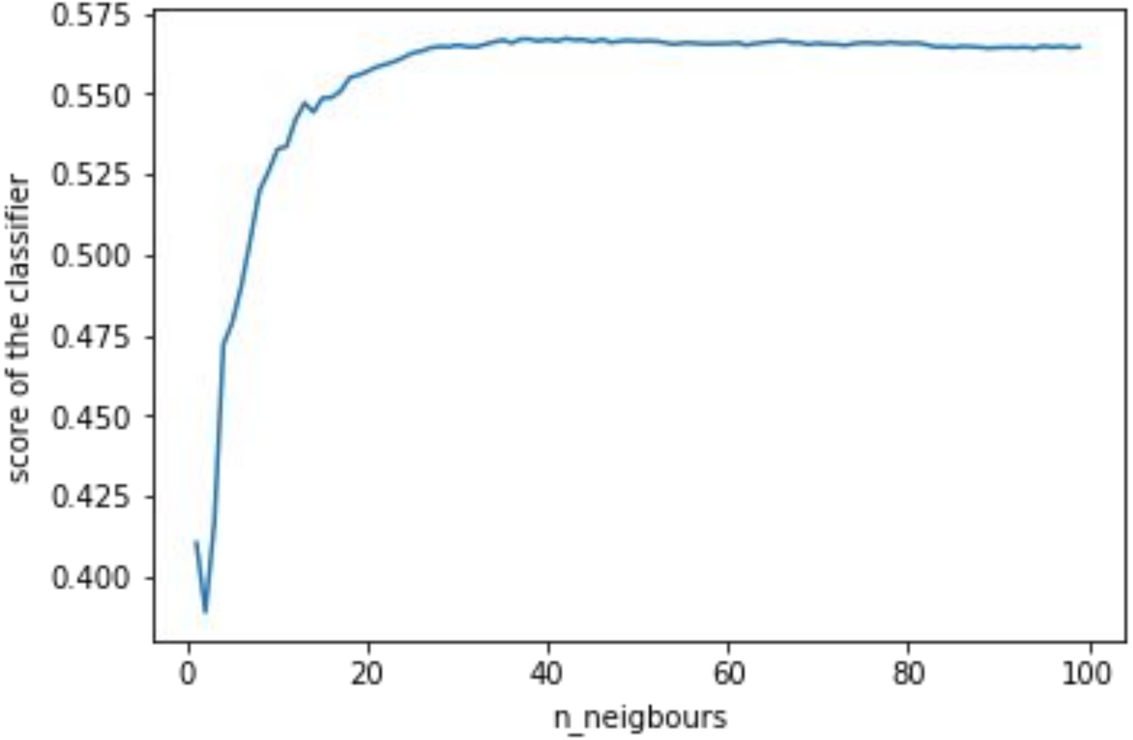
Step by Step Approach

1. Normalize the data
2. Create the training set and test set
3. Build the classifier using KNeighborsClassifier from sklearn.neighbors with n_neighbors=5
4. Fit the data into the classifier
5. Check the performance of the classifier using .score() to return the mean accuracy on the given test data and labels
6. Optimize the performance of the classifier with different values of n_neighbors

Classification of Education Levels

- We model education levels using income and the average length of essays.
- With n_neighbors=5, the score of the classifier is 0.479.
- The accuracy we would expect from predicting a education level by randomly selecting one would be 1/4, or 0.25.
- Our model outperform this number.
- The chart in the next slide depicts the progression of the accuracy of the classifier with increasing values of n_neigbors.

Classification of Education Levels using income and average length of essays



The optimal score is 0.567,
when n_neigbors is around 42.

Classification of Education Levels using other features

- We experiment with different combinations of features and the summary of the results are as follows:

Features used	Optimal Scores	n_neigbors
income, essay_ave_len	0.567	42
do_smoke, do_drug	0.54	5
do_smoke, do_drug, income, essay_ave_len	0.543	83
income, age	0.604	69

Conclusion

- We conclude that additional features such as do_smoke and do_drug do not improve the score of the classifier.
- The best score that we have obtained in our experiment is 0.604 where the feature used to learn the model is **income and age**. The accuracy we would expect from predicting a education level by randomly selecting one would be 1/4, or 0.25.
- Our model outperform the accuracy of randomly selecting an education level.

Support Vector Machines

- We build models using Support Vector Machine approach to classify gender using some combinations of height, income, age and do_smoke.
- An SVM makes classifications by defining a decision boundary and then seeing what side of the boundary an unclassified point falls on.
- Slide 4 indicates that a decision boundary can be made to classify gender. For example, if the user's height is above 65, we can guess that the user is male.

Step by Step Approach

1. Remove data with NA values.

2.Create the training set and validation set.

3. Build the classifier using SVC(kernel='rbf').

4. Fit the training data into the classifier.

5. Check the performance of the classifier using .score() to return the mean accuracy on the given test data and labels.

6. Optimize the performance of the classifier with different values of gamma and C.

Classification of Gender

- We model classifier for gender using heights.
- The core of the classifier is 0.836.
- The score is considered very high, and hence we can conclude that the classifier is quite accurate at predicting gender using heights.
- We experiment with different values of gamma and C values and find that the optimal score is 0.836, when gamma=4 and c=1.

Classification of Gender using different features

We experiment with different combinations of features and the summary of the results are as follows:

Features used	Optimal Scores
height, do_smoke	0.829
height, income	0.837
height, age	0.836

Our model of predicting gender with height and income gives score of 0.837 which significantly outperform the accuracy of randomly selecting a gender, which is 0.5.

Regression Approaches

Multiple Linear Regression

- We performed multiple linear regression to predict income using features such as gender_code, height, education_level, drinks_code, job_status_code, age, essay_ave_len, state_code.
- We explore the correlation of different features with income, and obtained the results tabulated in the next slide (features are sorted in decreasing level of correlation with income).

Correlation of Income with different features

Feature	Correlation with Income
gender_code	-0.076
height	0.067
education_level	-0.053
drinks_code	0.043
essay_ave_len	0.0084
job_status_code	0.0069
state_code	-0.0037
age	-0.0007

decreasing absolute correlation

Step by Step Approach

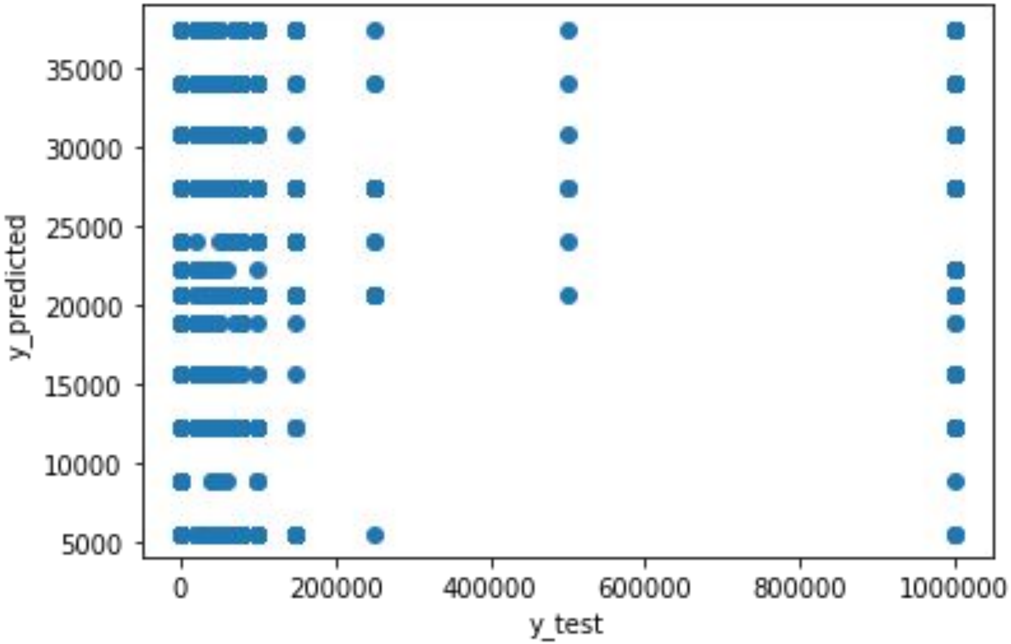
1. Separate out features to model on.
2. Remove data with NA values
3. Create the training set and testing set.
4. Build the LinearRegression model, and fit the training data into model.
5. Evaluate the performance of the model using .score().
6. Calculate the predicted incomes for our testing data and compare them to their actual incomes.

Feature: gender_code and education_level

- The score for the model is 0.00768.
- The score is very low, and indicates that the features used did not able to perform linear regression on income with satisfying accuracy.
- The coefficients on our different features :
  ('gender_code', -15141.210630850219)
  ('education_level', -3343.895436252767)
- This indicates that gender has higher impact than education level in the linear regression model on income using these features.

Feature: gender_code and education_level

We calculate the predicted incomes for our testing data and compare them to their actual incomes.



For a perfect linear regression model we would expect to see the data plotted along the line y = x, indicating homoscedasticity. In this case, it is not. We call this model heteroscedastic.

Other features

We define different subsets of data to fit into our model:

1. binary_features
2. numeric_features
3. all of the features
4. The top three highest correlated features
5. Features with correlation such that the absolute value greater than 0.01.

Results

Features Subset	Score
binary features	0.0406
numeric features	0.0115
all features	0.0115
top three highest correlated features with income	0.0086
features with correlation such that the absolute value greater than 0.01.	0.011

It appears that all of our regression model has very weak score.
We can conclude that our features are not sufficient to model income with reasonable accuracy.

Next Steps /Conclusions

Conclusion

- Performance on classification of heights and educations levels are satisfactory.
- We can predict a person’s height using gender and income.
- We can predict a person’s education level using income and age.
- However, performance on regression analysis of income is extremely weak given the selection of features that we used.

Next Step

- We should attempt to explore different combinations of features that can increase the performance of regression model on income.
- One potential feature that we can experiment is the job data. Can the job description of a user’s profile says something about the income?

Appendices

The following documents (available in the format of .ipynb and .html) contain the details codes of the models and plots that we used to derive results in this presentation:

- Education_Level_Classification
- Gender_Classification
- Income_Regression