Background
EMSI, the project’s sponsor company, currently use a machine learning algorithm to classify online job postings to gain data on the labor market. Our goal was to improve the model they currently use to find educational requirements in postings.

Objectives
The objective of this project is to find a classification system for job postings, categorizing them based on educational requirements: high school diploma or GED, associate’s degree, bachelor’s degree, master’s degree, and/or PhD.

Solution Approach
Our solution approach was primarily based around experimentation. First, we experimented with different classification architectures, implementing each model using Keras with TensorFlow as a backend. Once we found an ideal model, we tried different pre-processing techniques, and modified and refined the model until we achieved the results shown below.

Pre-Processing
Successful approaches:
- Removing non-alpha characters/stop words
- Stemming words (closely, closing -> close)
- Data augmentation: created extra PhD examples

Unsuccessful approaches:
- Recursive feature elimination
- Concentration metric -> Eliminate words with similar probability to appear in each class

Analysis
The main conclusion of our research in this classification task is that the most suitable classifier for the task, among the solutions we tested, was the CNN, which was both faster and more accurate than the alternative models. We also found that data augmentation succeeded in increasing the classification accuracy for Associate, Master and PhD postings, but caused overfitting on Bachelor posts.

Future Work
- Implement the model in a lighter framework to increase training efficiency
- Experiment with ensembling
- Modification of the convolutional neural network to improve accuracy
- Implement additional normalization and regularization techniques

Glossary
- GRU: Gated Recurrent Unit
- LSTM: Long Short Term Memory
- CNN: Convolutional Neural network

Results
Here are best results for the CNN using cross validation and data augmentation.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS/GED</td>
<td>0.94</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>Associate’s</td>
<td>0.84</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Bachelor’s</td>
<td>0.74</td>
<td>0.92</td>
<td>0.82</td>
</tr>
<tr>
<td>Master’s</td>
<td>0.86</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>PHD</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Acknowledgements
A special thanks to Justin Whear for mentoring this project, Professor Crandall for everything that he does, and the EMSI intern for spending 20 hrs per week hand classifying job postings.