“Project Alien Worker”
DATASCI W210 Capstone

Lucas Dan
Samuel Kabue
Sarah Neff
The Business Problem

● How long does it take to get a work visa for a given job title?

● Context and Audience
  ○ Companies want to know if it’s worth the effort (time, money) to fill post with a foreign worker
  ○ Individuals want to know what job areas have a quicker path to visa approval

● Impact
  ○ Improved insight on visa petitions on time taken to process applications for a given job type

● Importance
  ○ Companies can do more predictable planning about their hiring practices
  ○ Individuals can focus their applications to jobs that have better visa turnaround times
The Data

- **Collection and Extraction**
  - H-1B data collected 2001-2015 from Department of Labor
  - Merged into a single SQL database table with a combined schema

- **Cleanup**
  - Standardized labels (e.g. wage rates, state names), addresses, etc; trimmed whitespace
  - Created job dimension tables for modeling and higher-level analyses
  - Other pre-aggregated datasets to improve Tableau performance

- **Sample fields**
  - Application submission date
  - Job title
  - Last significant event (Certified, Denied, Withdrawn)
  - Last significant date (e.g. date certified)
The Solution

- **Process**
  - For each application, calculate “Days Until Visa Certified” variable
  - Use a sample set to train a KNN machine learning algorithm with job title as predictor
  - Then use the algorithm to predict “Days Until Visa Certified” for other inputs

- **Explanation**
  - Exploratory analysis shows job title can be a predictor for how long an application might take

- **Our differentiator**
  - Currently, official USCIS site only lets users/companies view current application stage
  - There are no sites or apps that try to predict the application time
  - Our project’s prediction adds an important piece that’s currently missing
The Data Pipeline

Data Extraction → H1B & PERM Data → Data Extraction

Raw Text Files → Data Processing

Processed Text Files → Web Application

Tableau Web Server

Interactive Visualizations (View #1)
- Calculate “Days Until Visa Certified” Variable

Interactive Visualizations (View #2)
- Company Suggestions
- Model Inputs

Relational Database

SQL Server

Data Storage

Python

Data Processing

Tableau

High-Level Statistics & Visualizations

Machine Learning Algorithm (KNN to Predict Days)

Individualized Prediction for User
Exploration: Application Status by Company

Top 10 Companies: Application Status

Employer Name:

Application Status:
- CERTIFIED
- CERTIFIED-WITHDRAWN
- DENIED
- WITHDRAWN

Number of Records:
- INFOSYS LIMITED: 89,000
- ACCENTURE LLP
- DELOITE CONSULT.
- ERNST & YOUNG
- HCL AMERIC.
- IBM INDIA PRIVATE
- LARSEN & TOUBRO
- MICROSOFT
- TATA CON.
- WIPRO LIMITED
Exploration: Days to Process by Company

Top 10 Companies: Days of Visa Processing
Exploration: Quickest Job Titles and Companies

**Quickest Certification Rate: Jobs**

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPUTER PROGRAMMER</td>
<td>1,073.87</td>
</tr>
<tr>
<td>PROGRAMMER ANALYST</td>
<td>1,117.77</td>
</tr>
<tr>
<td>SYSTEMS ANALYST</td>
<td>1,126.94</td>
</tr>
<tr>
<td>COMPUTER SYSTEMS ANALYST</td>
<td>1,132.24</td>
</tr>
<tr>
<td>SENIOR SOFTWARE ENGINEER</td>
<td>1,136.51</td>
</tr>
<tr>
<td>BUSINESS ANALYST</td>
<td>1,140.52</td>
</tr>
<tr>
<td>TECHNOLOGY LEAD - US</td>
<td>1,145.56</td>
</tr>
<tr>
<td>SOFTWARE DEVELOPER</td>
<td>1,152.83</td>
</tr>
<tr>
<td>SOFTWARE ENGINEER</td>
<td>1,156.11</td>
</tr>
<tr>
<td>TECHNOLOGY ANALYST - U</td>
<td>1,163.99</td>
</tr>
</tbody>
</table>

**Quickest Certification Rate: Companies**

<table>
<thead>
<tr>
<th>Employer Name</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM INDIA PRIVATE LIMITED</td>
<td>971.8</td>
</tr>
<tr>
<td>TATA CONSULTANCY SERVICES LLP</td>
<td>1,019.6</td>
</tr>
<tr>
<td>LARSEN &amp; TOUBRO INFOTECH LIMITED</td>
<td>1,054.5</td>
</tr>
<tr>
<td>HCL AMERICA, INC.</td>
<td>1,131.9</td>
</tr>
<tr>
<td>WIPRO LIMITED</td>
<td>1,145.0</td>
</tr>
<tr>
<td>DELOITTE CONSULTING LLP</td>
<td>1,149.8</td>
</tr>
<tr>
<td>INFOSYS LIMITED</td>
<td>1,163.2</td>
</tr>
<tr>
<td>ACCENTURE LLP</td>
<td>1,165.2</td>
</tr>
<tr>
<td>ERNST &amp; YOUNG U.S. LLP</td>
<td>1,179.2</td>
</tr>
<tr>
<td>MICROSOFT CORPORATION</td>
<td>1,192.2</td>
</tr>
</tbody>
</table>
The Predictive Modeling

- **Dependent Variable**
  - Days Until Visa Application Certified

- **Independent Variables**
  - Desired company location
  - Desired job title

- **K-Nearest Neighbors algorithm**
  - Predicts days until application is certified
  - Outputs 3 company suggestions with minimal wait times
  - User can input company location or job title for prediction

- **Accuracy**
  - Predictions on average are roughly within 50 days of true value
  - ~95% of predictions are within 200 days of true value
Highlights, Challenges, Takeaways

- **Project management from day 1**
  - Agile methodology allowed us to iterate over ideas and assigned tasks

- **Clear problem-solving approach, avoiding data overload**
  - Datasets had millions of rows and hundreds of variables
  - Great to explore all of the data, but should be quick to define problem statement

- **Dirty data causes problems**
  - Wrong or missing data and unstandardized fields
  - As a data originator, good to define a data structure/schema that holds with time
The Future

- Improving our training model by incorporating more predictors
  - Year of application
  - Applicant’s level of education
  - Applicant’s current visa class

- Incorporate additional datasets
  - E.g. if unemployment rates in a region affect visa turnaround time
The Project Website

http://people.ischool.berkeley.edu/~samkabue/capstone/
The Toolkit