What's in the Job Title?

Sylvia Lee

Thinkful Capstone Project

Problem

- Gap between actual salary and job seekers' expectation
- Job seekers spend time on researching salaries

Solution

- Allow job seekers to estimate salary based on job titles


Job Title
Location
Years of Experience
JOB
INCOME

Purpose of Study

- Build an algorithm to estimate annual income based on target job title, employment length, and location
- Examine the importance of job title in predicting income

Data

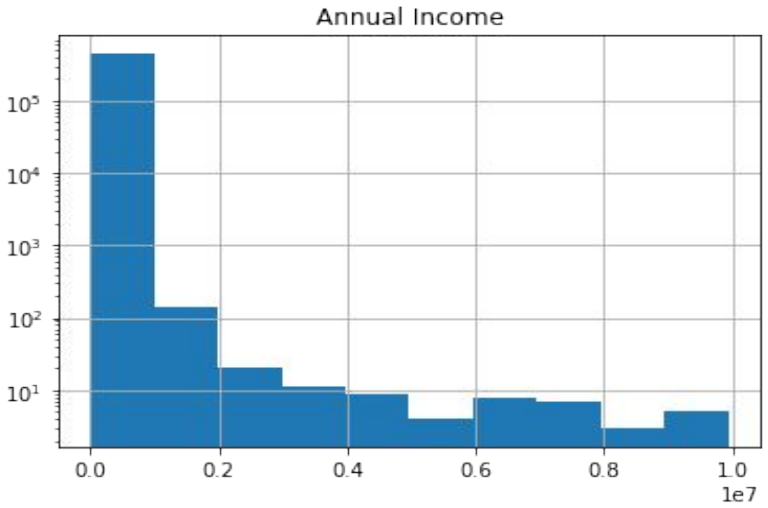
- Subset from Lending Club loan dataset
- 2018
- 495,242 records
- 4 fields
    - 2 numeric variables: annual income, employment length
    - 1 geographic variable: zip code
    - 1 natural language variable: job title

Exploratory Data Analysis

- Drop missing values
  - job title (11%)
  - employment length (8%)
- 440,555 records

Target Variable

- Exclude records with salary less than 2.5K

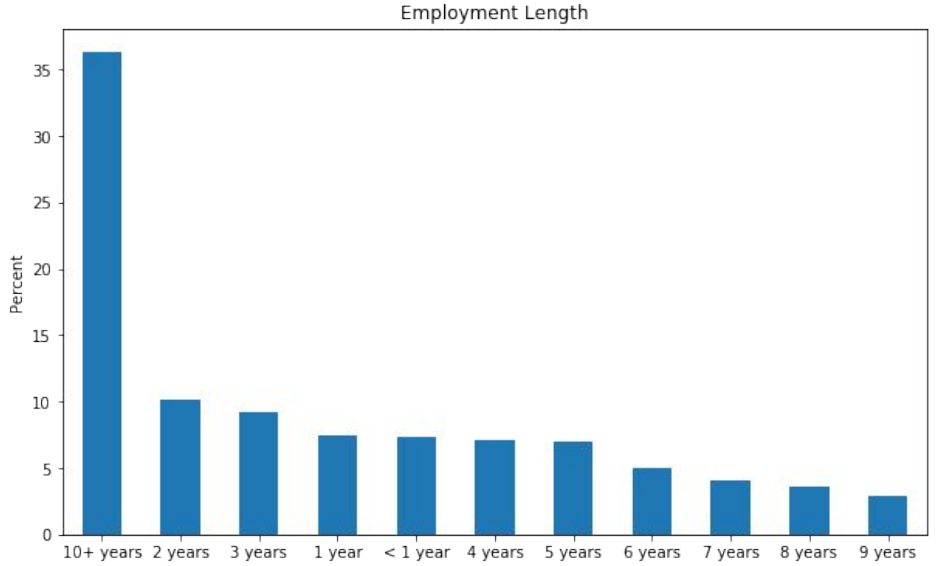


Predictors

Predictor	Example	Type	Cardinality
Job Title	Chef	Natural Language	129,443
Zip Code	109xx	Categorical	895
Employment Length	10+ years	Categorical	11

Feature Engineering

- Employment Length
  - Categorical → Ordinal
- Zip Code
  - One-hot encoding
- Job Title
  - NLP: abbreviations
  - spaCy: En_core_web_lg
  - 300 dimensional vectors




```
#300 dimensions
job18_vect.head()
```

0	1	2	3	4	5	6	7	8	9	10	11
-0.190910	-0.286760	0.325610	-0.815440	0.50561	0.789190	0.788340	-0.39848	-0.017570	1.94030	-0.313080	-0.182890
0.234640	0.698470	-0.400190	-0.652680	-0.04554	0.095587	0.021045	-0.91542	-0.186420	0.97901	-0.154840	-0.180760
-0.148010	-0.413780	0.523630	0.240100	-0.41241	0.078526	0.185500	-0.15445	0.137410	2.89810	-0.503730	-0.251790
-0.053464	-0.026130	-0.014067	-0.029245	0.08011	0.243942	0.007125	-0.46131	-0.184965	2.62500	-0.156605	0.053418
0.074737	-0.009538	-0.354550	-0.007188	0.21565	-0.327910	-0.574090	-0.35732	-0.416150	1.49760	0.162470	0.172090

Final Dataset

```
X = pd.get_dummies(job_feats, drop_first=True).values
y = job_final['annual_inc'].values

print(X.shape)
print(y.shape)
```

```
(440419, 1195)
(440419,)
```

Build Models

- Stochastic Gradient Descent
- Ridge Regression
- Lasso Regression
- Neural Network (Deep Learning)
    - 2-5 Layers

Hyperparameters

- Alpha
- Loss function
- Eta0
- Learning Rate
- L1_Ratio
- PCA zip code

Final Model

Model	Layer	Epochs	Loss	R-squared Train	R-squared Test	RMSE Train	RMSE Test
Neural Network	2	100	MSE	0.16	0.14	$81K	$78K

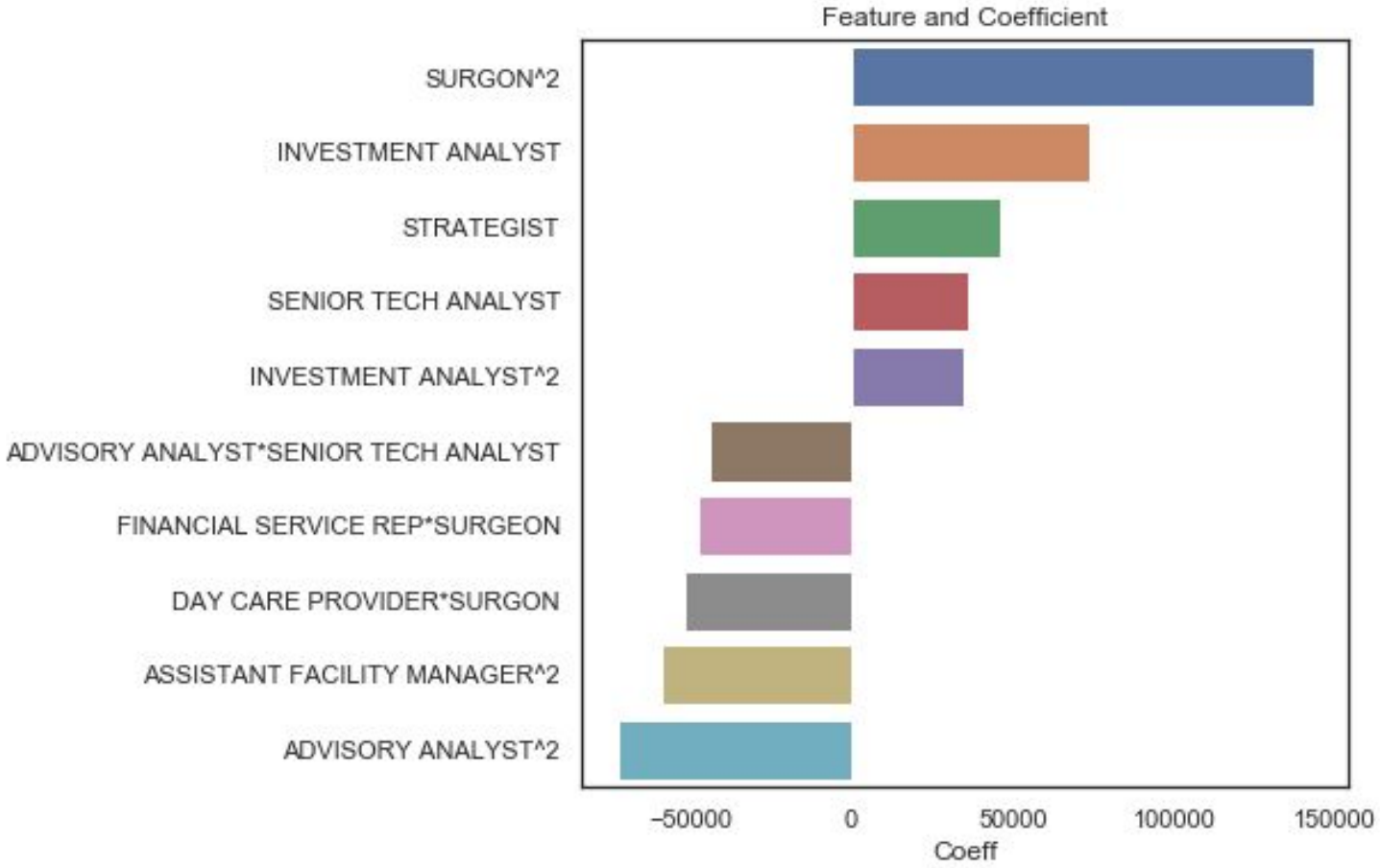
Semantic Analysis

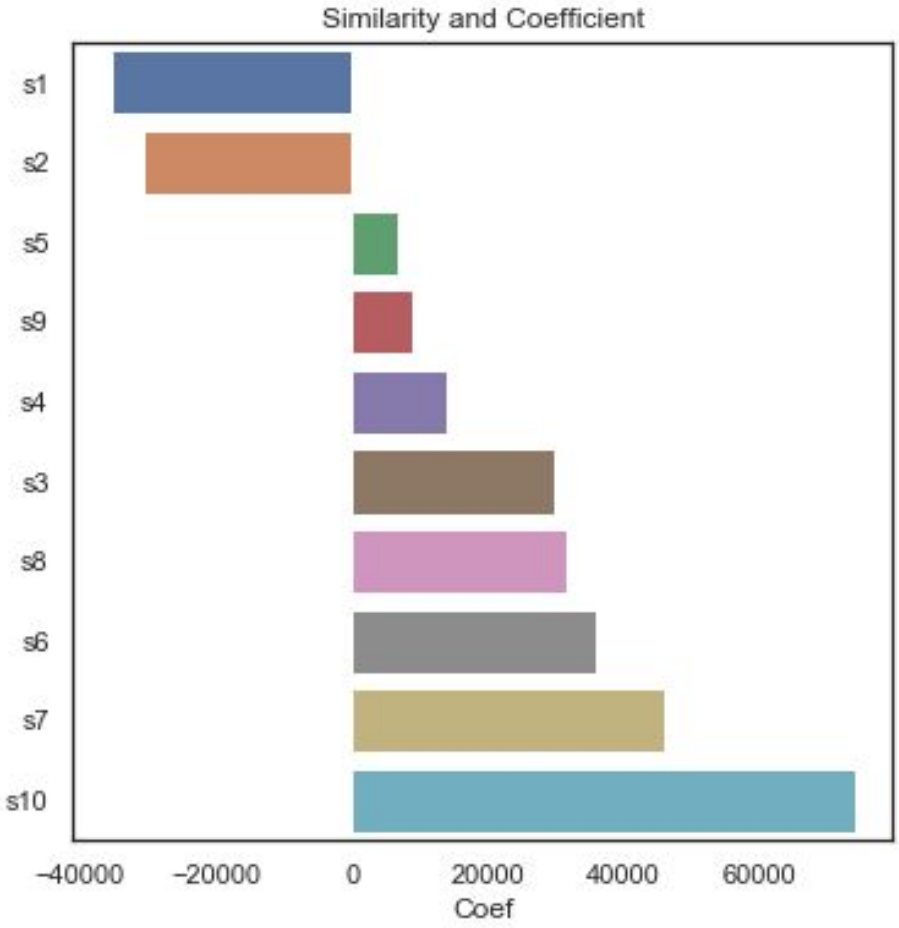
- Use only job title to predict salary for zip code 112xx
  - Sort data by annual income
  - Segment data based on quartiles
  - Select two job titles to represent each group as reference
  - Compute cosine similarity on the 300d word2vec vectors to determine the similarity between each job title and each reference job title
  - Use 10 similarity scores as features to predict annual income

INCOME	JOB TITLE
Top	INVESTMENT ANALYST
	SURGEON
75%ile	MANAGEMENT CONSULTANT
	STRATEGIST
50%ile	SENIOR TECH ANALYST
	ASSISTANT FACILITY MANAGER
25%ile	FINANCIAL SERVICE REP
	ADVISORY ANALYST
Bottom	BOOKSELLER
	DAY CARE PROVIDER

Final Model

Model	R-squared Train	R-squared Test	RMSE Train	RMSE Test
Polynomial Ridge Regression	0.13	0.12	$55k	$48k


Feature and Coefficient
SURGON^2
INVESTMENT ANALYST
STRATEGIST
SENIOR TECH ANALYST
INVESTMENT ANALYST^2
ADVISORY ANALYST*SENIOR TECH ANALYST
FINANCIAL SERVICE REP*SURGEON
DAY CARE PROVIDER*SURGON
ASSISTANT FACILITY MANAGER^2
ADVISORY ANALYST^2
−50000
0
50000
100000
150000
Coeff


Similarity and Coefficient
s1
s2
s5
s9
s4
s3
s8
s6
s7
s10
−40000
−20000
0
20000
40000
60000
Coef

Future Work

- Add time series data
- Find more data for each zip code
- Incorporate numeric geographic variables if possible
- Chose those titles more rigorously by looking for the most common title in each of the 5 levels in the income range