Census Income Level Prediction

Group 6

Wan Chen, Zhe Li, Gabriel Roth,
Tiffany Yang, Horace Zhen

Agenda

01 Motivation

02 Data Summary

03 Results

04 Features Interpretation

05 Model Comparison

06 Conclusion

Motivation

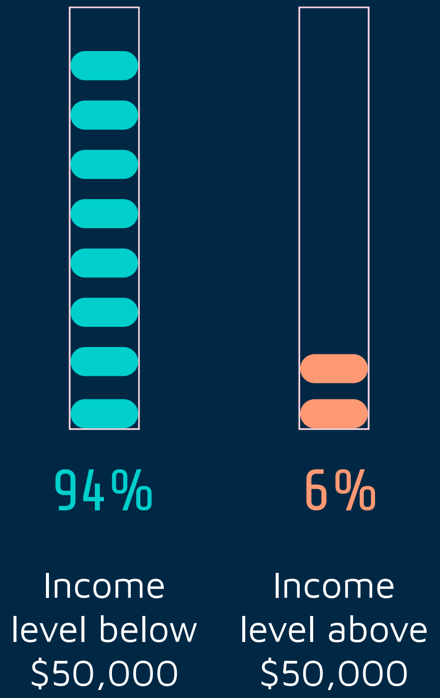
01

- Understand the US demographics
- Classify if personal income level is below or above $50k
- Why binary classification task with cut-off at $50k?
    - 200% of the personal average income
    - 9 times the poverty line
    - Separates the top 24% and the rest
- Understand what are the factors that are keeping people from making to the top 24%

Data Summary

02

- File size: 156 MB
- Total number of rows: 299,285
- Total input features: 41
    - Numeric variables: 8
    - Categorical variables: 33
- Binary label
    - Below $50,000
    - Above $50,000



Categorical Variables
education-related
work-related
demographics-related
employment-related
residence-related
family-related
veteran status

Numeric Variables
age
wage per hour
capital gains
capital losses
dividends from stocks
instance weight
number of persons worked for employer
number of weeks worked in a year

Data Preparation

Missing Values
Filled with "unspecified" in an additional category

Instance Weight
Dropped for during classification

Encode the Label
Class 0: < $50k
Class 1: > $50k

Feature Selection
Fitted the training set to a Random Forest Classifier and retained features (importance > 0.005)

Results

03

Classification Models

01 Bernoulli Naive Bayes

02 Mixed Naive Bayes

03 Logistic Regression

04 Random Forest

Results

Accuracy	Bernoulli Naive Bayes	Mixed Naive Bayes	Logistic Regression	Random Forest
Without Feature Selection	0.74	0.86	0.95	0.95
With Feature Selection	0.83	0.90	0.95	0.95

F1-Score	Bernoulli Naive Bayes	Mixed Naive Bayes	Logistic Regression	Random Forest
Without Feature Selection	0.30	0.43	0.51	0.52
With Feature Selection	0.38	0.48	0.48	0.53

Recall	Bernoulli Naive Bayes	Mixed Naive Bayes	Logistic Regression	Random Forest
Without Feature Selection	0.90	0.83	0.40	0.40
With Feature Selection	0.82	0.77	0.36	0.45

Precision	Bernoulli Naive Bayes	Mixed Naive Bayes	Logistic Regression	Random Forest
Without Feature Selection	0.18	0.29	0.73	0.73
With Feature Selection	0.25	0.35	0.72	0.66

Specificity	Bernoulli Naive Bayes	Mixed Naive Bayes	Logistic Regression	Random Forest
Without Feature Selection	0.73	0.87	0.99	0.99
With Feature Selection	0.83	0.90	0.99	0.98

Features Interpretation

04

Why are some features important?

age	0.090709
dividends_from_stocks	0.073793
capital_gains	0.066063
num_persons_worked_for_employer	0.039914
weeks_worked_in_year	0.028934
capital_losses	0.020817
education_ Bachelors degree(BA AB BS)	0.016674
sex_ Female	0.016287
major_occupation_code_ Executive admin and managerial	0.016254
education_ Masters degree(MA MS MEng MEd MSW MBA)	0.016240
sex_ Male	0.015678
occupation_code_2	0.015100
education_ Prof school degree (MD DDS DVM LLB JD)	0.013858
detailed_household_summary_in_household_ Householder	0.013809
major_occupation_code_ Professional specialty	0.012799
education_ High school graduate	0.011182
tax_filer_status_ Joint both under 65	0.009522
education_ Some college but no degree	0.009293
wage_per_hour	0.009016
education_ Doctorate degree(PhD EdD)	0.008779
class_of_worker_ Private	0.008658
own_business_or_self_employed_0	0.008194
member_of_a_labor_union_ Not in universe	0.008107
member_of_a_labor_union_ No	0.007296
own_business_or_self_employed_2	0.006970
detailed_household_summary_in_household_ Spouse of householder	0.006798
class_of_worker_ Self-employed-incorporated	0.006776
marital_status_ Married-civilian spouse present	0.006649
major_industry_code_ Not in universe or children	0.006170
marital_status_ Never married	0.006149
occupation_code_4	0.005733
full_or_part_time_employment_stat_ Full-time schedules	0.005527
country_of_birth_father_ United-States	0.005325
race_ White	0.005185
occupation_code_7	0.005163
major_occupation_code_ Not in universe	0.005079
class_of_worker_ Self-employed-not incorporated	0.005029

```
lasso_features

[(0.06966793889950086, 'major_occupation_code_Executive admin and managerial'),
 (0.04555635443397677, 'major_occupation_code_Professional specialty'),
 (0.005088009528619055, 'num_persons_worked_for_employer'),
 (0.004755082330769386, 'detailed_household_summary_in_household_Householder'),
 (0.0035502481794541948, 'tax_filer_status_Joint both under 65'),
 (0.003281872552894073, 'education_Bachelors degree(BA AB BS)'),
 (0.0015108166263765495, 'weeks_worked_in_year'),
 (0.0007401410445909137, 'age'),
 (9.691033351989998e-05, 'capital_losses'),
 (1.5859210647824184e-05, 'divdends_from_stocks'),
 (9.769573895177861e-06, 'capital_gains'),
 (2.968241918663065e-17, 'sex_Male'),
 (-1.6196313980702442e-05, 'wage_per_hour'),
 (-0.015202162430726668, 'tax_filer_status_Single'),
 (-0.025885379794678937, 'education_High school graduate'),
 (-0.03661256139220707, 'sex_Female')]
```

Why are some features important?

- We are unable to interpret the full 472 feature set.
- Selected top 37 features for modeling and interpretation:
- Top features with positive importance: age, dividends from stocks, capital gains, capital loss, self-employed status, weeks worked in year, education - bachelors and above, sex male, occupation
- Top features with negative importance: sex female, education - high school, tax status - single, wage per hour
- Important aspects for interpretation include age (time), education and career choice, work status, demographic factors, and sexuality

Model Comparison

05

Why is a certain method better than other methods?

Random Forest

- Ensemble model: averages the predictions of many decision trees
- Train DT on bootstrapped samples and random subsets of features -> diversity

Logistic Regression

- Data can naturally be represented as a linear combination of factors
- For example: weeks worked, capital gains/losses, and education, all contribute to a higher income
- Sigmoid function adds nonlinearity; the model is able to capture the relationship between the input features

Why is a certain method better than other methods?

Bernoulli Naive Bayes

- Large number of variables with many categories -> needs a lot of data to get accurate likelihood probabilities
- Naive Bayes assumption is not realistic for census data
- An individual's job position is tied to factors such as their education, age (experience), etc.

Mixed Naive Bayes

- Capable of mixing CategoricalNB and GaussianNB from scikit-learn -> Better than BernoulliNB

- The discriminative models were better on all metrics besides recall since they learn which features are important for classification. However, since the dataset is imbalanced the recall is lower.

Conclusion

06

What is the extracted knowledge from this data?

- The formula of making it to the top income level is a combination of working hard, generating multiple income streams, investing in capitals, and investing early.
  - The impact of wage per hour on income is almost natural with an slight negative impact.
  - Capital gains, capital loss, and dividend from stocks all have a high importance score and impact the outcome positively.
  - The age (time) effect - with the power of compound interest, investment will grow significantly with a long time horizon.
- Gender inequality
  - Females are negatively correlated with the outcome, implying gender inequality exists in the workplace, including unequal pay and disparity in promotions.

References

[1] Sharath R., Krishna Chaitanya S., Nirupam K. N., Sowmya B. J., and K. G. Srinivasa, "Data analytics to predict the income and economic hierarchy on Census data," 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 249-254, 2016, doi: 10.1109/CSITSS.2016.7779366.

[2] Chakrabarty, N., & Biswas, S., "A statistical approach to adult census income level prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 207-212, 2018. doi: 10.1109/ICACCCN.2018.8748528.

[3] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277–292, 2010. doi: 10.1007/s10618-010-0190-x.

[4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," Machine Learning and Knowledge Discovery in Databases, pp. 35–50, 2012. doi: 10.1007/978-3-642-33486-3_3.

[5] Ding, F., Hardt, M., Miller, J., & Schmidt, L., "Retiring adult: New datasets for fair machine learning, ArXiv, abs/2108.04884. doi: 10.48550/arXiv.2108.04884.

[6] A. B. Atkinson and A. Brandolini, "On the identification of the middle class," Income Inequality, pp. 77–100, 2013.

Thank You!