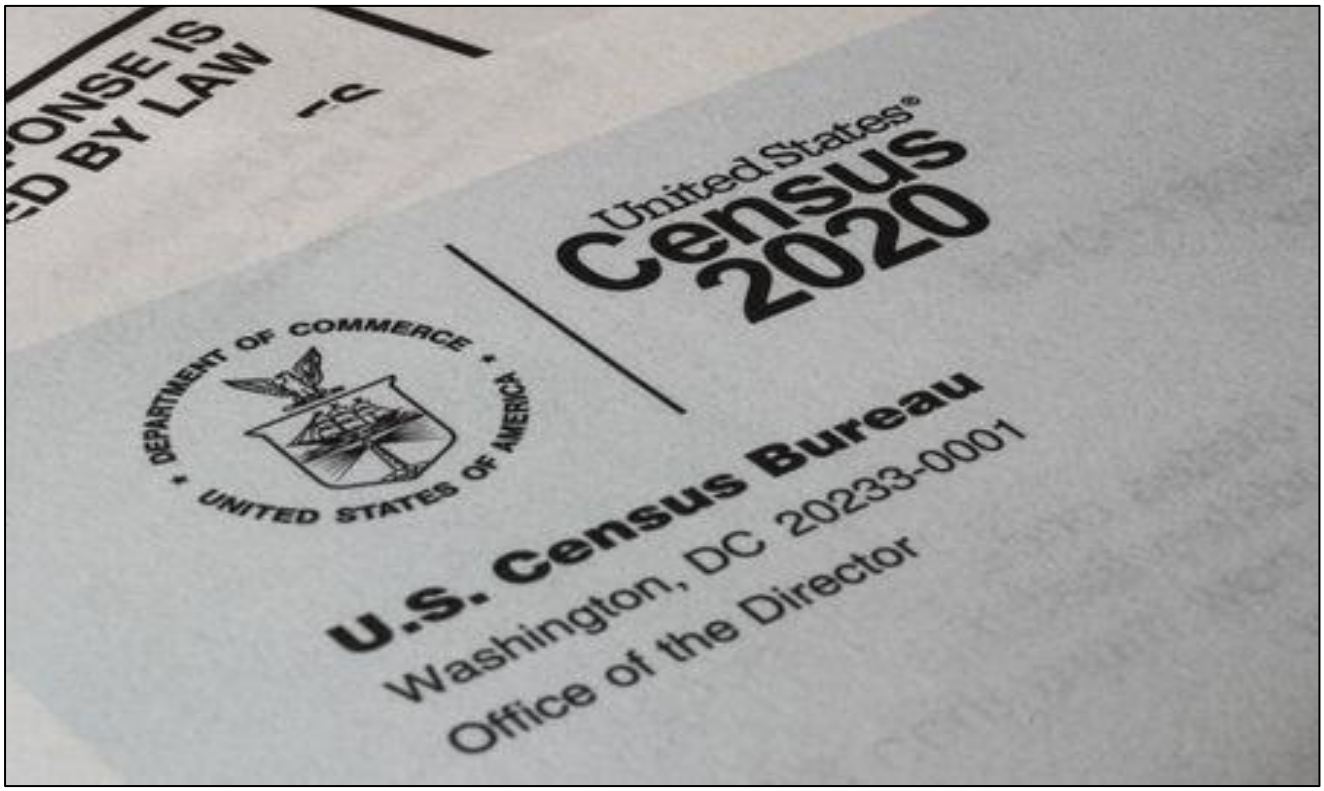
CENSUS INCOME PREDICTION MODEL



Submitted by:

UJJWAL PRATIK

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT CENSUS INCOME

Census money income is defined as income received on a regular basis before payments for taxes, social security, etc. and does not reflect noncash benefits.

ABOUS CENSUS BUREAU

The Census Bureau's mission is to serve as the nation’s leading provider of quality data about its people and economy. Their Goal is to provide the best mix of timeliness, relevancy, quality and cost for the data collect and services provide.

PROBLEM STATEMENT

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). The prediction task is to determine whether a person makes over $50K a year.us.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Age
- Work class
- Final weight
- Education
- Education num
- Marital Status
- Occupation
- Relationship
- Race
- Sex
- Capital gain.
- Capital Loss
- Hours Per Week
- Native Country
- Income (Target Variable)

sample

Workclass	Fnlwgt	Education	Education_num	Marital_status	Occupation	Relationship	Race	Sex	Capital_gain	Capital_loss	Hours_per_week	Native_cour
Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-Sta
Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-Sta
Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-Sta
Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-Sta
Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-Sta

OBJECTIVE

- Our main objective is to prediction of task to determine whether a person makes over $50K a year.
- All the parameters will be analysed through Machine Learning algorithms like Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier etc which will help to predict whether a person makes over $50K a year.

DATA DESCRIPTION

- The source of data is taken from GitHub.
- Data Link is (https://raw.githubusercontent.com/dsrscientist/dataset1/master/census_)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the census income because it gives the best outcome.
- The independent variable is Age, Work class, Final weight, Education, Education num, Occupation, Relationship, Race, Sex, Capital gain, capital loss, Hours per week, Native Country, and dependent Variable is Income.

METRIC USAGE

a. Logistic Regression.
b. AdaBoost Classifier.
c. Random Forest Classifier.
d. Gaussian NB
e. Decision Tree Classifier.
f. Support Vector Classifier

SYSTEM REQUIREMENTS

HARDWARE and SOFTWARE REQUIREMENTS and TOOLS USED

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- AdaBoost Classifier: s a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
- Random Forest Classifier: The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.
- Gaussian NB: Gaussian NB is based on the Naive Bayes theorem with the assumption of conditional independence between every pair of features given the label of the target class.
- Decision Tree Classifier: A decision tree is a flowchart-like tree structure in which the internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. A Decision Tree consists of Nodes: Test for the value of a certain attribute.
- Support Vector Classifier: Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- ZScore: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- IMB Learn Anaconda: imbalanced-learn is a python package offering a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contra projects.
- Standard Scaling: Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.
- AUC-ROC CURVE: The AUC-ROC curve tells us visualize how well our machine learning classifier is carrying out. It is one of the most significant evaluation metrics for examining any classification model's performance. It is also called as AUROC (Area Under the Receiver Operating Characteristics).

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Label Encoding
c. STANDARD SCLAER (SMOTE)
d. IMB Learn
e. Logistic Regression
f. AdaBoost Classifier
g. Gaussian NB
h. Decision Tree classifier
i. Support Vector Classifier
j. Grid Search CV
k. Cross Validation
l. AUC ROC CURVE

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Analysed data for distribution.
- Compared between two columns.
- Checked and removed outliers through zscore method.
- Removed skewness present in the dataset.
- Done Oversampling.
- Done Standard Scaling.
- Cross validate the accuracy score from overfitting.
- Done AUC ROC score for better understanding.
- Hyper Parameter tuning using Grid Search CV.

CONCLUSION

As our conclusion we proclaim that, after checking accuracy score, cross validation, Ensemble Techniques, and checking AUC score we declare Random Forest Classifier is predicting 100% accuracy is best suited model for our purpose of predicting whether a person makes over $50K a year.

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.