Developing Marketing Profiles using Exploratory Data Analysis and Predictive Analytics

Shashi Kiran Chilukuri
Arizona State University
Schiluk6@asu.edu

James Graves
Arizona State University
jwgrave2@asu.edu

Kenneth Han
Arizona State University
khan23@asu.edu

Yinghai Zhao
Arizona State University
yzhao310@asu.edu

***Abstract*—Data visualization is an essential component of Exploratory Data Analysis (EDA). It represents data in the graphical format. It uses various visual elements like charts, graphs, maps, trees etc. to provide an accessible way to see and understand trends, patterns, outliers, and relationships between different variables in the data. In this project, we used these techniques along with predictive analytics to identify most important factors that influence an individuals income range, and developed predictive model and marketing profiles. To perform these tasks, we used python and its libraries. This paper will first introduce to the project, provide the solution in detail, followed by the results, personal contribution, lessons learned and references.**

***Keywords*—Heatmap, Bar plot, Box plot, Mosaic plot, Donut plot, Correlation, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine(SVM)**

I. Introduction

As part of this *CSE 578: Data Visualization* course's team project, need to use data visualization techniques along with predictive analytics to develop marketing profiles, identify the factors that determines individual's income class and to predict whether the income of an individual is above or below $50K. To achieve this, we were given with data supplied by the US Census Bureau and were asked to focus on $50K as a key number for salary. And to develop the solution, we divided the project into two phases, namely, "Exploratory Data Analysis" and "Predictive Analytics" phases.

In the phase one of the project, explored the data using various data visualizations to find the trends, patterns, relationship between the variables etc., and checked for any interesting inferences we may get. For that, we converted categorical data into numerical form and performed correlation between all the variables to identify highly correlated variables. Upon these identified variables, we have used various plots to find interesting inferences. At the end, our goal was to develop marketing profiles based on these relevant inferences.

In the phase two of the project, performed data preprocessing steps such as separating features and label, converting the categorical variables to 1's and 0's, feature scaling and splitting the dataset into train and test datasets. Once the data is ready for predictive models, implemented four machine learning models, namely, Logistic Regression, Decision Tree, Random Forest and Support Vector Machine (Linear) to identify important factors that determine important factors and to predict whether the income of an individual is above or below $50K threshold.

Dataset information:

- Data Source: United States Census Bureau
- Features: 14 Features (8 Categorical and 6 continuous variables)
- Label: Income with 2 Categories (">50K","<=50K")
- Initial Dataset Length: 32,561 records
- Data Cleaning: Removed records with unknown ("?") data, Stripped white spaces
- Final Dataset Length: 30,162 records

Resources used in this project:

	Resources
Programming language	Python
Integrated Development Environment (IDE)	Jupyter Notebook, Python Environment
Python Libraries	numpy, pandas, matplotlib, seaborn, SKlearn
Visualizations	Bar, Stacked bar, Box and whisker, Scatter, Mosaic, Pie, Heatmap
Models/Methods	Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (Linear), Standard Scaler, train$_test_split$

II. Solution

Phase One: Exploratory Data Analysis

In this phase, to explore the data and to identify underlying patterns and relationships, first converted the categorical variables into numbers and then plotted the correlation heatmap (Fig.1) between them. Now, by applying the correlation threshold of >0.20 and <-0.18 on the correlation values obtained from heatmap, identified important features that are highly (positively/negatively) correlated with the label "Income". Here is the list of important features:

- Education-num
- Age
- Sex
- Capital gain
- Hours per week
- Relationship
- Marital status

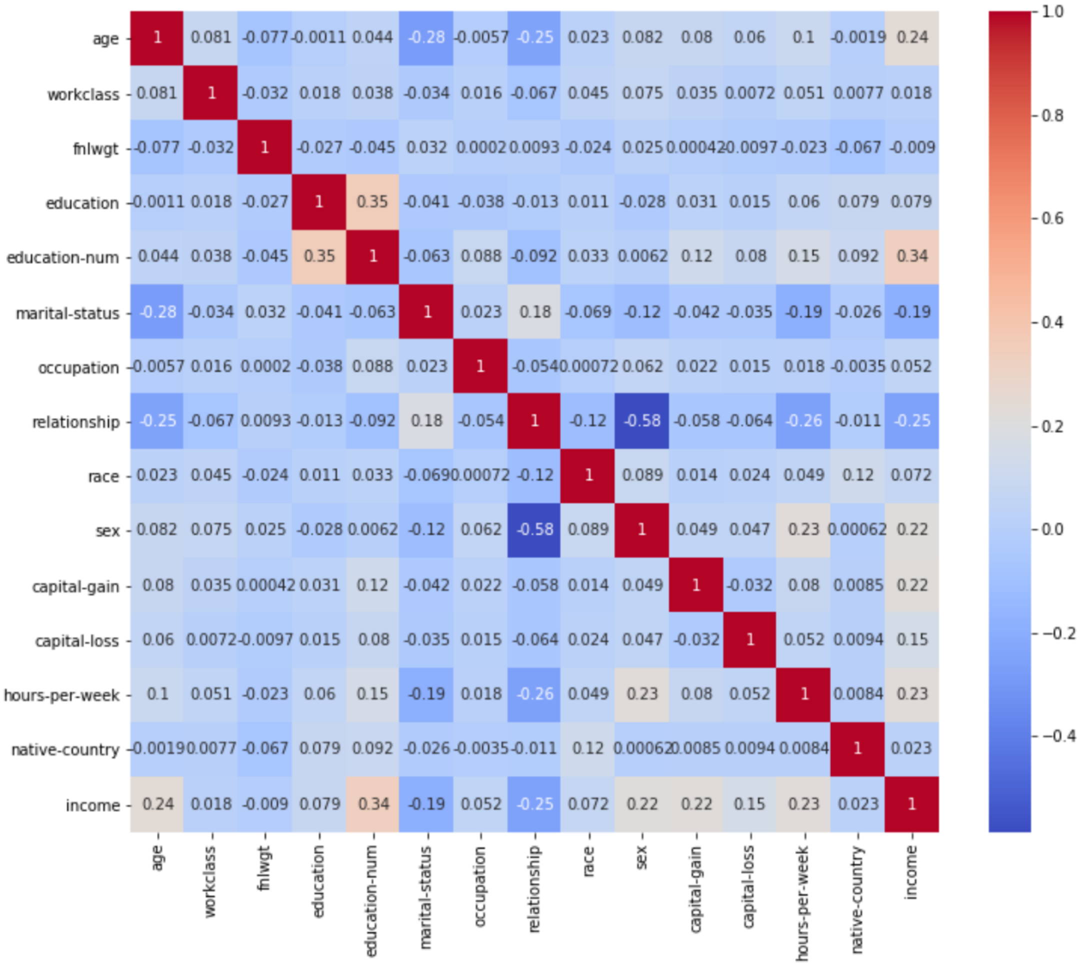


Fig. 1. Data Correlation Heatmap

To further explore these important features, visualized each one of them with the class label.

1) Plots between Education-num vs Income: Plot1 is a Bar plot of feature "Education-num" with two income bins on X axis and count on Y axis, and Plot2 is Box plot between Label "income" and Feature "Education num" on X and Y axis respectively.

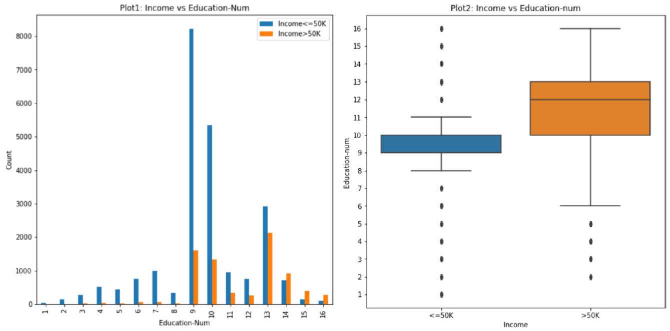


Fig. 2. Plot between Age vs Income

Here are the inferences from these plots:

- From Plot1: Individuals with education more than 13 years have higher chances making more than $50K.
- From Plot2: 75% of the group making <=50K have less than 10th grade education as opposed to only 25% of >50K income group.
- From Plot2: For the <=50K income group, 12 years of education is considered as an outlier.

2) Plot between Income vs Age: It is a bar plot between Label "income" and Feature "age" on X and Y axis respectively.

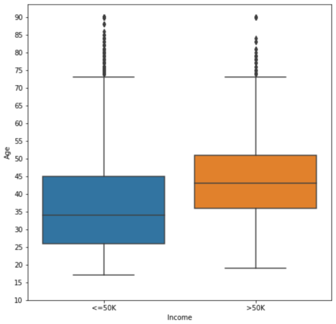


Fig. 3. Plots between Income vs Age

Here are the inferences from this plot:

- Individuals who are making income >50K are older than those making <=50K at the Quartile1, Quartile2 (median) and Quartile3.
- Those making <=50K have a greater Inter Quartile Range (IQR) which means greater group diversity with respect to age.
- Box plot establishes ages exceeding 74 are outliers.

3) Plot between Income vs Sex: It is a mosaic plot between Label "income" and Feature "sex" on X and Y axis respectively.

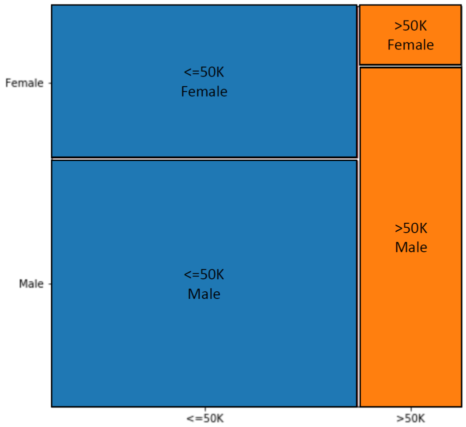


Fig. 4. Plots between Income vs Sex

Here is the inference from this plot:

- There is a high probability that individual making more than 50K is a Male.

4) Plots between Hours-per-week vs Income: Plot1 is a Stacked Bar plot between Feature "Hours-per-week" with two income ranges on X-axis and with count on Y axis, and Plot2 is Box plot between Label "income" and Feature "Hours-per-week" on X and Y axis respectively.

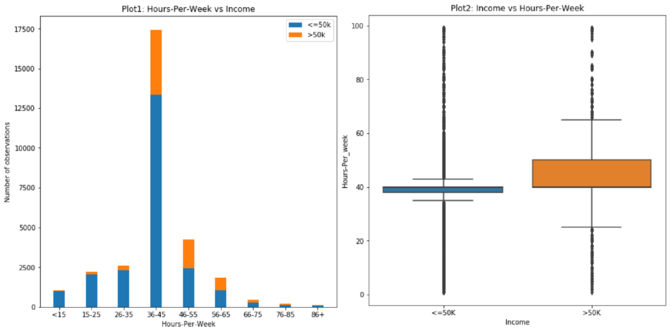


Fig. 5. Plots between Hours-per-week vs Income

Here is the inference from this plot:

- From Plot1: Individuals making more than 50K are likely to be working more than 35 hours per week.
- From Plot2: Inter Quartile range is small for the <=50K group which indicates a high concentration of hours worked near 40 hour work week. In case of over 50K income group, IQR is larger and, Q1, Q2(median) and Q3 are all higher than the respective values in the <=50K income group.

5) Plot between Relationship vs Income: It is a donut plot between Feature "Relationship" and Label "income".

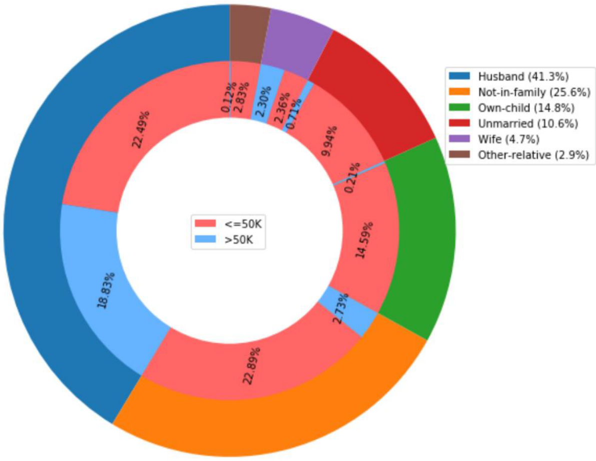


Fig. 6. Plots between Relationship vs Income

Here is the inference from this plot:

- Individual with Relationship category "Husband" or "Wife" has a higher probability of making more than 50K as compared to categories "Not in family", "Own child", "Unmarried" and "Other relative".

6) Plot between Income vs Marital-staus: It is a mosaic plot between Label "income" and Feature "Marital-Status" on X and Y axis respectively.

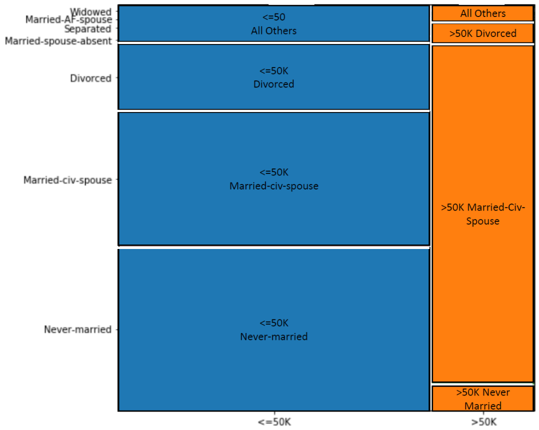


Fig. 7. Plots between Income vs Marital-Status

Here is the inference from this plot:

- Individual making more than 50K very likely to have marital status "Married-civ-spouse".

Finally, here is list of six interesting inferences that we got from exploratory data analysis of important features with the label "income":

- For the <=50K income group, 12 years of education is considered as an outlier.
- The median ages for the <=50K income group is 34 whereas for the >50K income group it is 44.
- There is a high probability that individual making more than 50K is a Male.
- Individuals making more than 50K are likely to be working more than 35 hours per week.
- Individual with Relationship category “Husband” or “Wife” has a higher probability of making more than 50K as compared to categories “Not in family”, “Own child”, “Unmarried” and “Other relative”.
- Individual making more than 50K very likely to have marital status “Married-civ-spouse”.

Phase Two: Predictive Analytics

In this phase, implemented four machine learning models, namely, Logistic regression, Decision Tree, Random forest and Support Vector Machine (Liner) models to identify the important factors that influenced the income class of an individual and predicted whether an individual’s income is above or below $50K. To implement these models, performed following data pre-processing steps:

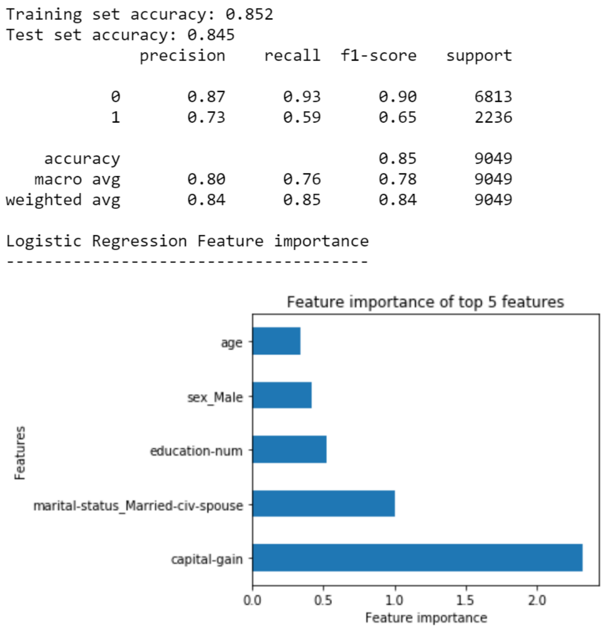
- Separated dataset into Features (with 14 variables) and Label ”Income” sets.
- Identified 8 categorical features and converted them into 1’s and 0’s using Pandas ”get dummies” method resulting in a total of 96 different features.
- For the model to give equal importance to all the features, performed featuring scaling and normalized the data.
- Performed 70/30 dataset split to get training and testing sets respectively. As a result following train and test sets are created:

Dataset	***Sample size***
Feature Train Set	21113 X 96
Label Train Set	21113 X 1
Feature Test Set	9039 X 96
Label Test Set	9039 X 1

Results of four Models:

1) Logistic Regression Implementation: This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”. Here are results from this model:

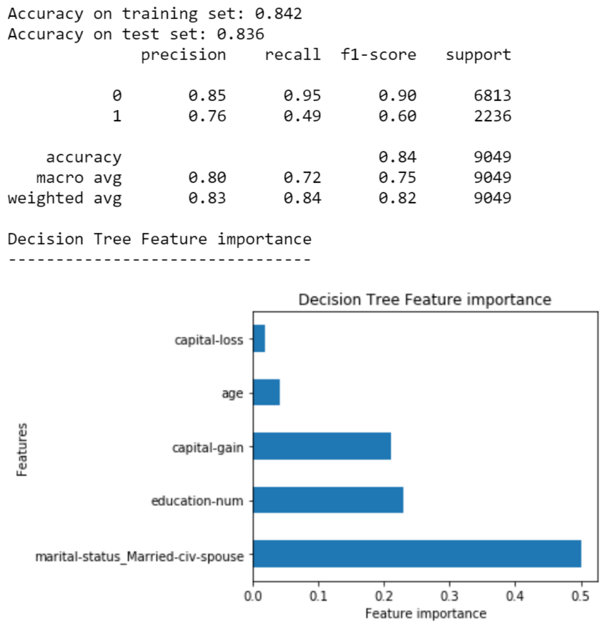
- This model classified the test data with 84.5% accuracy.
- Top 5 Important features suggested by this model are:
  1) Captial-gain
  2) Martial-status”Married-civ-spouse”
  3) Education-num
  4) Sex”Male”
  5) Age



```
Training set accuracy: 0.852
Test set accuracy: 0.845
              precision    recall  f1-score   support

           0       0.87      0.93      0.90      6813
           1       0.73      0.59      0.65      2236

    accuracy                           0.85      9049
   macro avg       0.80      0.76      0.78      9049
weighted avg       0.84      0.85      0.84      9049

Logistic Regression Feature importance
--------------------------------------
```

Fig. 8. Logistic Regression Model Results

2) Decision Tree Implementation: This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”. To identify best parameters for the model, performed grid search.Based on the search results, it is identified that model works best with criterion=’entropy’ and max depth=4. Used these parameters for model implementation. Here are results from this model:



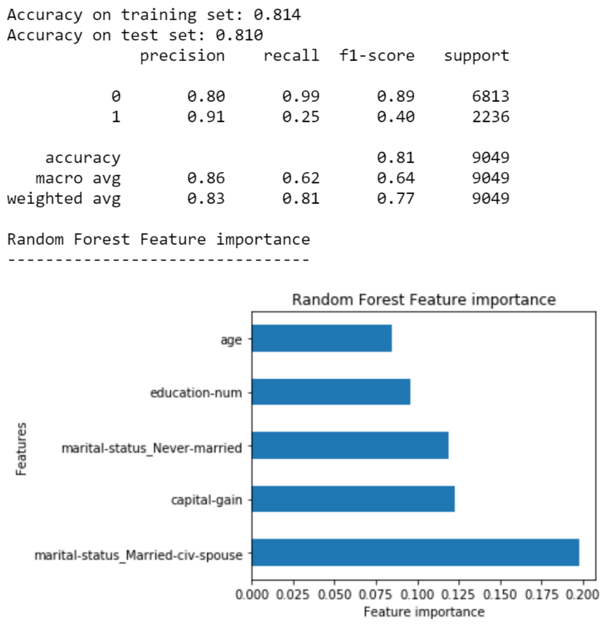
```
Accuracy on training set: 0.842
Accuracy on test set: 0.836
              precision    recall  f1-score   support

           0       0.85      0.95      0.90      6813
           1       0.76      0.49      0.60      2236

    accuracy                           0.84      9049
   macro avg       0.80      0.72      0.75      9049
weighted avg       0.83      0.84      0.82      9049

Decision Tree Feature importance
--------------------------------
```

Fig. 9. Decision Tree Model Results

- This model classified the test data with 83.6% accuracy.
- Top 5 Important features suggested by this model are:

1) Martial-status”Married-civ-spouse”
2) Education-num
3) Captial-gain
4) Age
5) Sex”Male”

3) Random Forest Implementation: This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”. To identify best parameters for the model, performed grid search. Based on the search results, it is identified that model works best with criterion=’entropy’ and max depth=4. Used these parameters for model implementation. Here are results from this model:



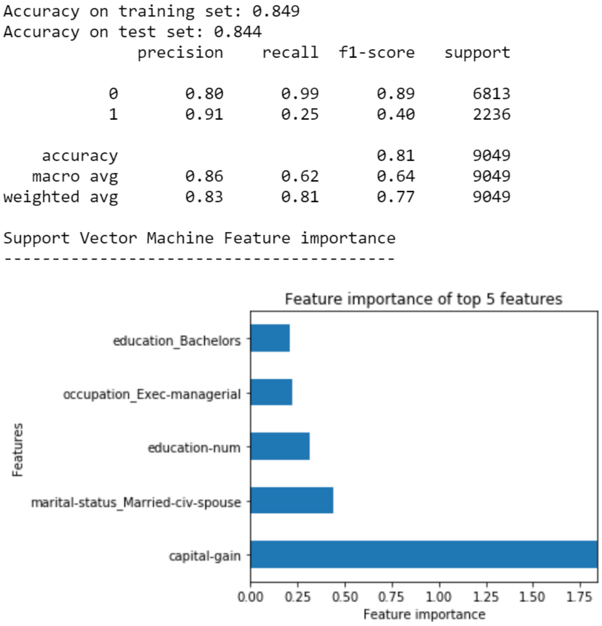
```
Accuracy on training set: 0.814
Accuracy on test set: 0.810
              precision    recall  f1-score   support

           0       0.80      0.99      0.89      6813
           1       0.91      0.25      0.40      2236

    accuracy                           0.81      9049
   macro avg       0.86      0.62      0.64      9049
weighted avg       0.83      0.81      0.77      9049

Random Forest Feature importance
--------------------------------
```

Fig. 10. Random Forest Model Results

- This model classified the test data with 83.6% accuracy.
- Top 5 Important features suggested by this model are:
  1) Martial-status”Married-civ-spouse”
  2) Captial-gain
  3) Martial-status”Never-Married”
  4) Education-num
  5) Age

4) Support Vector Machine Implementation: This model will classify given features into two class categories “Income more than 50K” and “Income less than or equal to 50K”. Used ”Linear” kernal for this model. Here are model results:

- This model classified the test data with 84.4% accuracy.
- Top 5 Important features suggested by this model are:
  1) Captial-gain
  2) Martial-status”Married-civ-spouse”
  3) Education-num
  4) occupation ”Exec-managerial”
  5) education ”Bachelors”

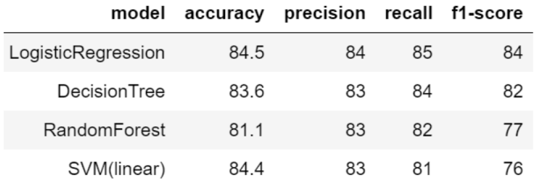


```
Accuracy on training set: 0.849
Accuracy on test set: 0.844
              precision    recall  f1-score   support

           0       0.80      0.99      0.89      6813
           1       0.91      0.25      0.40      2236

    accuracy                           0.81      9049
   macro avg       0.86      0.62      0.64      9049
weighted avg       0.83      0.81      0.77      9049

Support Vector Machine Feature importance
-----------------------------------------
```

Fig. 11. SVM Model Results

III. Results and Recommendations

All four models implemented in this project have showed similar accuracy with Logistic Regression having a slight edge in predicting the results. Here is the model evaluation matrix:



model	accuracy	precision	recall	f1-score
LogisticRegression	84.5	84	85	84
DecisionTree	83.6	83	84	82
RandomForest	81.1	83	82	77
SVM(linear)	84.4	83	81	76

Fig. 12. Model Evaluation Matrix

We can visualize above results to identify the model to choose for our prediction. Here is the plot:

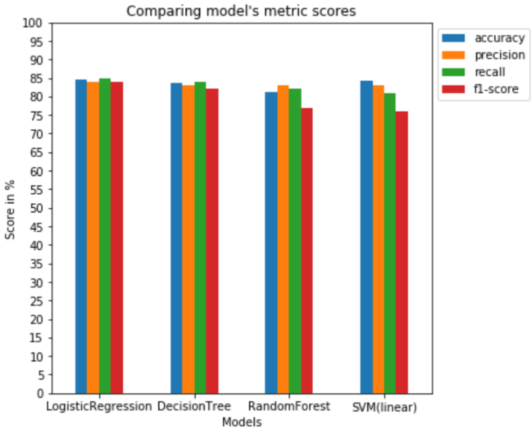


Fig. 13. Plot comparing Machine Learning Model results

Machine Learning Model Recommendations

- Based on our study of four popular predictive analytics models, we recommend the logistic regression model for predicting whether the income of an individual is above or below 50K and here is the ranked list of the important features suggested by the model:
    1) Captial-gain
    2) Martial-status"Married-civ-spouse"
    3) Education-num
    4) Sex "Male"
    5) Age
- Based on our analysis, the following features were found to be significant across all the machine learning models and also, in our exploratory data analysis:
    - Martial-status"Married-civ-spouse"
    - Education-num
    - Captial-gain

IV. Contributions

This is a team project. I worked with James Graves, Kenneth Han and Yinghai Zhao. In this project, I was involved in all the phases, right from understanding the project requirements to preparing the report for project submission. Here is the list of activities I performed as part of this project:

- Understand the project requirements and its tasks^3.
- Plan the project with other team members based on project deadlines/milestone.
- Project communication including scheduling zoom calls, Chat sessions etc.
- Create and maintain project to-do list.
- Create and maintain private code repository for version control.
- Perform initial data analysis to understand and look for any data discrepancies including checking data statistics, null data, data column naming etc. and prepare the Jupyter notebook with the cleaned-up data.
- Perform exploratory data analysis to understand how the features are correlated with the label and lay out the project execution strategy.
- Perform the data pre-processing, and develop the machine learning algorithms to identify the important factors and predict the income class.
- Review / Suggest the visualizations developed by other team members.
- Involved in preparing project executive report and system documentation report.
- Project code and document submission.

V. Lessons Learned

Here are some of the lessons learnt from this project:

- Data visualization is important and it helps to make the inferences from the data.
- Essence of Bertin's Visual Variables and how it enables us to plot better visualizations.
- Use of uni-variate and multi-variate visualizations as per the requirements.
- For categorical variables where there is no ordinal relationship, it is better to convert them to 1's and 0's (One-Hot Encoding). If we force the ordinal relationship by using Ordinal encoding, machine learning model will assume the natural ordering and may result in poor performance.

References

[1] Dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

[2] CSE 575 Statistical Machine Leaning, Spring B 2020, ASU

[3] Project requirements: https://www.coursera.org/learn/cse578/supplement/B4Lrb/course-project-introduction