CSE 578: DATA VISUALIZATION

Systems Documentation Report

Abhay Joshi (1213203951)
Aditya Chayapathy (1213050538)
Jagdeesh Basavaraju (1213004713)

Roles and Responsibilities:

- **Team members:** Abhay Joshi, Aditya Chayapathy, Jagdeesh Basavaraju
- **Stakeholders:** UVW College
- **Product owners:** XYZ Corporation

As a part of the team of data analysts at XYZ Corporation, the project given is to develop an application for UVW College to identify the factors that determine the individual's income. UVW College is looking to bolster their enrollment. They have identified salary as a key to determine criteria for marketing their degree programs. This prediction will be used to tailor their marketing efforts of reaching out to the individuals. Data supplied by the United States Census Bureau will be used as a dataset to build the application and will be focusing on $50,000 as a key number for salary.

The primary tasks included the understanding of dataset, finding factors that determine the individual's income and build machine learning models using the factors identified to predict the income of an individual.

The tasks distribution is as follows:

- Every member of the team had to understand the dataset individually.
- The features or attributes were distributed equally among the team members to analyse individually against the class label (salary-range) and come up with their analysis results.
- Feature Engineering: Abhay Joshi worked on identifying the initial set of factors to consider for further analysis using analysis results from the previous task.

- Visualizations: Jagdeesh Basavaraju worked on multivariate analysis and plots for the factors identified in the previous task.
- Machine learning analysis: Aditya Chayapathy worked on building models to predict the individual's income based on the factors identified using Machine Learning.
- Reiteration of visualizations were done by all team members.
- Documentation and report writing were done by all team members.

Team Goals and Business Objective:

The aim of this project is to identify the patterns in the dataset by plotting visualizations to help find the factors that account for determining the income of an individual and present the same to the UVW executives. Next, based on the initial analysis, we would like to build machine learning models that accurately predict the income of the individuals. The UVW marketing team would then use the application and the analysis results to tailor their marketing efforts of reaching out to the individuals.

Assumptions:

- **Dataset is accurate and precise:** We are assuming that the dataset given to us is accurate and precise. It cannot have any erroneous elements and must convey the correct message without being misleading. Without understanding how the data will be consumed, ensuring accuracy and precision could be off-target or more costly than necessary.
- **Legitimacy and Validity:** We presume that the dataset is legitimate and validated. For example, in the given dataset, items such as gender are typically limited to a set of options and open answers are not permitted. Any answers other than these would not be considered valid or legitimate based on the dataset's requirement.
- **Timeliness and Relevance:** We assume the dataset is collected at the right moment in time. Data collected too soon or too late could misrepresent a situation and drive inaccurate decisions. For example, the salaries in the dataset if sourced during a recession, it doesn't justify the dataset and will result in inaccurate results.
- **Completeness and Comprehensiveness:** Incomplete data is as dangerous as inaccurate data. So we assume that the dataset is complete and there are no gaps in data collection which might possibly lead to a partial view of the overall picture

to be displayed. Without a complete picture of how operations are running, uninformed actions will occur. It is important to understand the complete set of requirements that constitute a comprehensive set of data to determine whether or not the requirements are being fulfilled.

- **Feature Selection:** We assume that the features which contribute the most towards class prediction can give us more interpretable patterns. So, the majority of the visualizations and analysis is done on the selected features which are unbiased.

User Stories:

- **Data cleaning and preprocessing:** The dataset consists of various inconsistent and missing data ("?") which was preprocessed and the dataset was made compatible to run the models.
- **Individual attribute analysis:** The 14 features in the dataset was distributed equally among all the members of the group and individual feature analysis was performed.
- **Feature Engineering:** Abhay did the task of identifying the initial set of features to consider for further analysis using individual attribute analysis.
- **Visualizations:** Jagdeesh worked on multivariate analysis and plots for the factors identified in the feature engineering process.
- **Machine Learning:** Aditya worked on building model to predict the individual's income based on the factors identified using Machine Learning.
- **Executive and Systems Report:** All three members of the team worked on executive and systems report.

Visualizations:

Initially, each of the features was individually analysed through the use of data exploration techniques to identify patterns. This was achieved by leveraging data visualization tools such as pie charts, bar charts, histograms and box plots. Also, other statistical methods such as mean, median and standard deviation were used, when appropriate, to understand the underlying distribution. Next, we strengthened our initial analysis through multivariate feature analysis using tools such as mosaic plot, parallel coordinate plot and scatter plot. With this, we identified a set of important features. This formed the basis for the machine learning analysis that involved feature engineering of

the important features. Next, we trained few machine learning models to identify the individual's income. From this, we recognised the feature importance and observed that they were on par with our initial analysis of the features.

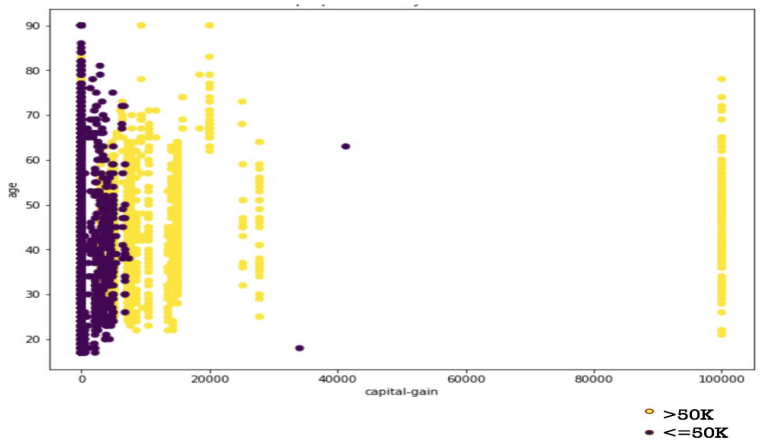
Initial analysis of individual features:

NOTE: Please find the graphs with respect to these inferences in the attached Jupyter notebooks. The graphs included in the report are restricted to the most influential plots.

1. **Age:** With increase in age the salary gets reduced very majorly. People in the age range of 30-55 have more chance of getting salary above 50k.
2. **Work class:** Not a great feature to consider since the data is very stable across both classes. Interesting to see more people who are in federal gov earn above 50k.
3. **Fnlwgt:** This attribute with values greater than 45k have salary <=50k.
4. **Education:** It is very evident from the graphs that people with education more than bachelors have higher chance of getting salary >50k.
5. **Education-num:** It is the same as education attribute.
6. **Marital status:** It's interesting to see that 85% of people earning more than 50K have Married-civ-spouse as marital status. It also shows that people who never married, earn less.
7. **Occupation:** It can be inferred that Exec-managerial and Prof-speciality occupation people make almost 50% of people earning more than 50K. Craft-repair occupation people are equally spread over the two classes.
8. **Relationship:** This particular feature has more influence on the salary as you can see people belonging to relationship "Husband" constitute 75% of the total people who are earning more than 50K. As identified by "Marital-status" feature, even this feature inferences that unmarried or never married people tend to earn less than 50K.
9. **Race**: "Race" attribute doesn't really influence the salary range among the people. In both the classes, majority of people belong to "White" race.
10. **Sex:** Male dominate both the classes but the domination is much stronger among salary range greater than 50K.
11. **Capital gain:** If the capital-gain is high, it is more like that the person earns more than 50K.
12. **Capital loss:** The capital-loss is more or less the same for both the classes of people. So this may not be a good factor that can be used to classify the data.
13. **Hours per week:** Most people working more than 40 hours a week belong to the ">50K" category.
14. **Native country:** The people belonging to the following nationality are more likely to earn less than 50K:
    Jamaica, Mexico, Puerto-Rico, Honduras, Colombia, Ecuador, Laos, Haiti, Portugal, Dominican-Republic, El-Salvador, Guatemala, Peru, Outlying-US, Trinidad & Tobago, Nicaragua, Vietnam.

The following are the important graphs that helped us in identifying the important features in the dataset:

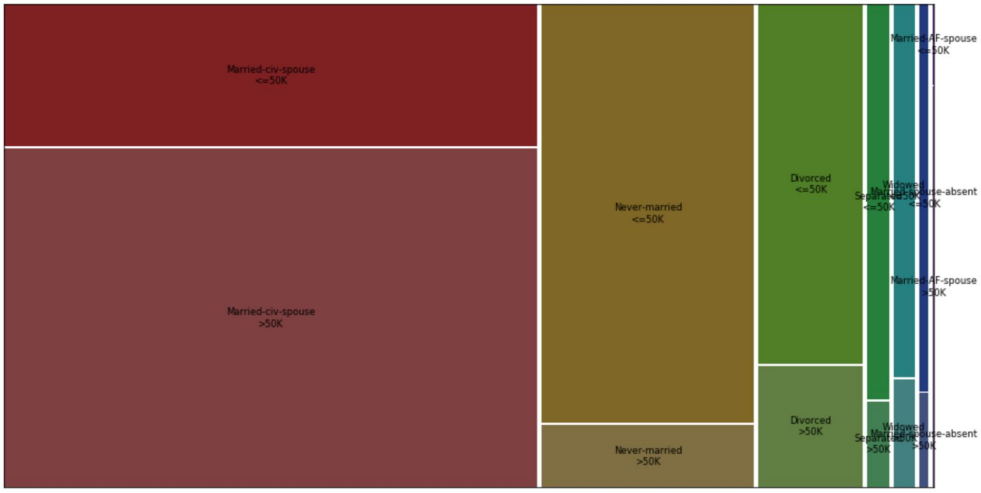
Scatter plot: Capital-gain and Age:



Inferences:

1. There seems to be a separation between the two classes of data with the exception of a few outliers.
2. Individuals with high capital gain are more likely to earn more than 50K income.

Mosaic plot: Marital status and Salary:



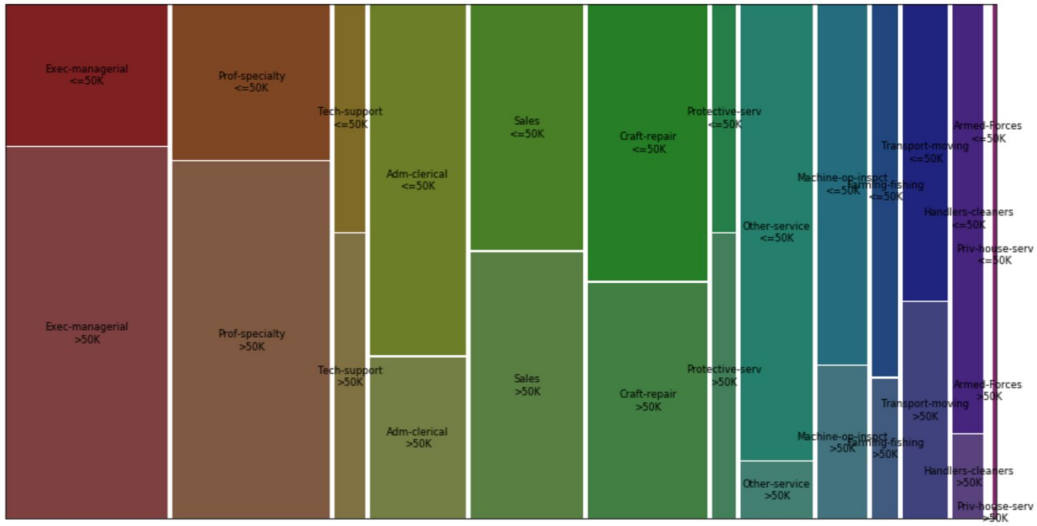
Categories in the order as they appear: Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed

Inferences:

1. For most categorical data, the distribution of the two classes are highly skewed hinting that this feature can be used to distinguish among the two classes.
2. Individuals with marital-status of “married-civ-spouse” are more likely to earn more than 50K income.

3. Individuals with marital-status of “never-married” are more likely to earn less than 50K income.

Mosaic plot: Occupation and salary:

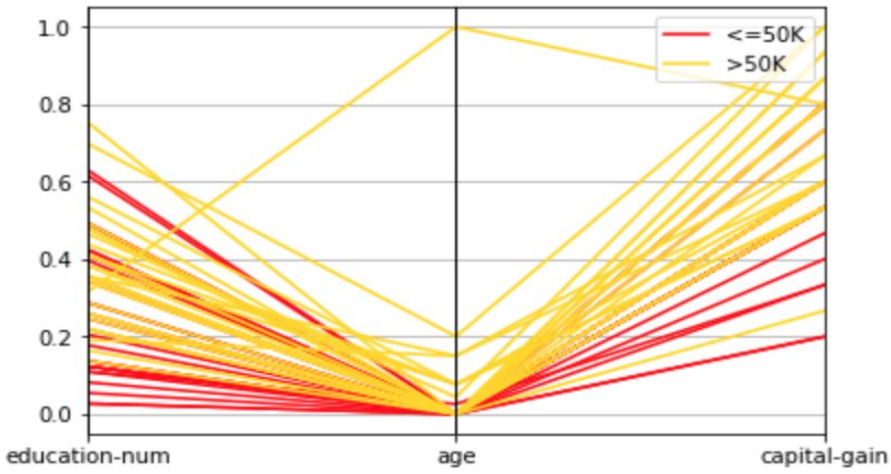


Categories in the order as they appear: Adm-clerical, Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Sales, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Craft-repair, Protective-serv, Armed-Forces, Priv-house-serv

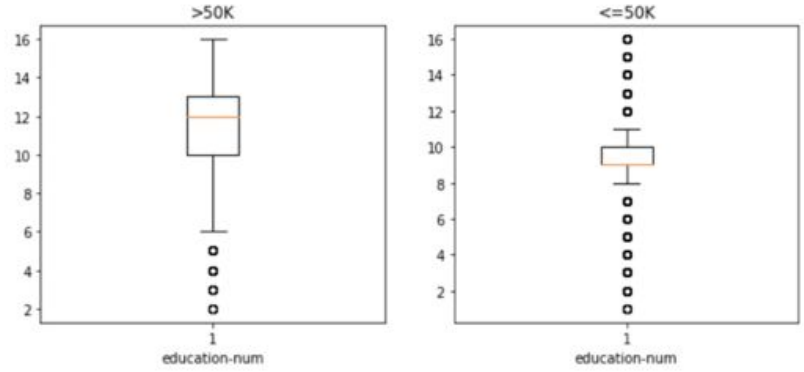
Inferences:

1. For most categorical data, the distribution of the two classes are highly skewed hinting that this feature can be used to distinguish among the two classes.
2. Individuals with occupations such as “Exec-managerial”, “Prof-speciality” are more likely to earn 50K income.

Parallel Coordinate Plot: Capital-gain, Education-num and Age



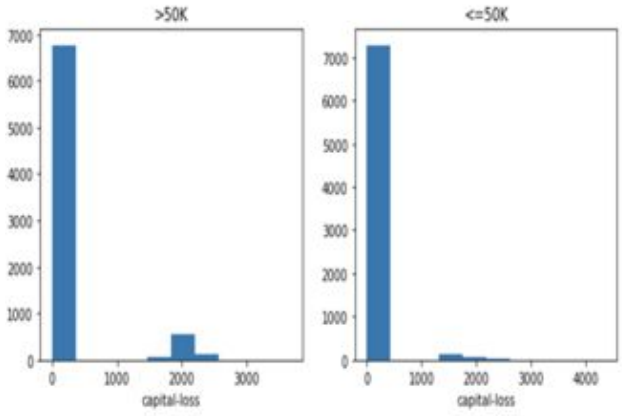
Box plot: education-num



Inferences:

1. From the parallel coordinate plot, we can see that the yellow lines and the red lines can be distinguished using the combination of these three features.
2. From the box plot, we can see that the distribution of the education among the two classes of data vary drastically.
3. Individuals with high education number are more likely to earn greater than 50K income.
4. Older individuals are likely to earn more than younger individuals.

Histogram: capital-loss



Inferences:

1. As seen in the figure, for both classes of data, they show similar distribution indicating that this feature may not help in distinguishing between the two classes of data.

Machine Learning Analysis:

NOTE: You can find the machine learning analysis in the attached Jupyter notebooks.

Steps Involved:

1. **Feature Engineering:** All the numerical features were left as is. For each categorical data, a numerical number is assigned based on the distinguishing factor of that category from our initial data exploration analysis.
2. **Data Normalization:** Each feature was scaled to a value between 0 and 1. This is done to ensure that the ML algorithms give equal importance to each of the features.
3. **Data Division:** Data was split in the ratio of 80:20 where 80 percent of the data is used for training and 20 percent of the data is used for testing.
4. **Training ML Models:** The following algorithms were used for training purposes of the ML models:
    a. Gradient Boosting
    b. Random Forest
    c. Support Vector Machine
    d. Neural Network
    e. Bagging
    f. Logistic Regression
    g. Decision Tree
    h. K Nearest Neighbours
5. **Hyperparameter Tuning:** The hyperparameters for each of these algorithms were modified to get the most accurate results.
6. **Evaluating Performance:** The trained models were evaluated using measures such as F1 score, recall, precision and accuracy on the test dataset. The following are the performance evaluation results:

.

	Gradient Boosting	Random Forest	SVM	Neural Network	Bagging	Logistic Regression	Decision Tree	K Nearest Neighbors
Accuracy	0.8218	0.8189	0.8093	0.812	0.8118	0.8044	0.7944	0.787
F1 Score	0.8297	0.8292	0.823	0.8205	0.8193	0.8136	0.8025	0.7922
Precision	0.7966	0.7868	0.7699	0.7872	0.7902	0.7793	0.7743	0.7758
Recall	0.8657	0.8764	0.8839	0.8568	0.8506	0.851	0.8328	0.8092

Questions

- *The dataset is highly skewed as the class distribution is highly uneven(2:1) in the training data. The question here is that how to use this data to infer and analyze the visualization?*
    - We used the methodology of k-fold cross validation of data to overcome this problem. k-fold cross validation is a resampling procedure used to evaluate machine learning models on a limited data sample. We have used a limited sample from the dataset so that the class distribution is balanced and in order to estimate how the model/analysis-algorithm is expected to perform in general when used to make predictions on data not used during the training of the model. We reiterated this process keeping the class distribution same and introduced stochasticity by using random sampling from the data distribution.

- *What features to use and how to assign weightage towards each feature in the dataset?*
    - There are 14 features in this dataset with 6 continuous features and the rest categorical features. Not all the features contribute towards class prediction. We individually analysed each of the features using data visualization tools to understand their distribution and rank their importance. The features that have more noticeable differences in the patterns for the two classes of data were given a higher rank.

- *How to convert categorical data to numerical data?*
    - In order to run the machine learning models, categorical data cannot be used to train the model. So, we converted the categorical data into numerical data by one-hot encoding that maps each of the unique categorical value to a number. Categories that can easily distinguish between the two classes were assigned a higher value.

- *What kind of visualizations to use to perform data analysis?*
    - For categorical data, visualizations like the pie chart and mosaic plots (multivariate) were used as these plots represent the categorical data well.
    - For continuous data, visualizations like the histogram, box plots, scatter plots and parallel coordinate plots were used.

- *How to go about machine learning analysis?*
    - After converting categorical data to continuous data, the next step was to split the dataset into training and testing sets. We made an 80:20 partition where 80 percent of the data was used for training and 20 percent of the data was used for testing purposes.
    - Next, we identified a set of algorithms with varying underlying assumptions for classification.
    - We modified the hyperparameters of the algorithms get the best performance of each of the algorithms
    - Lastly, the performance of each of the algorithms was evaluated against the test dataset.

Not Doing

- As a future scope we are planning to implement visual recommenders which makes use of a query builder to get the features of a new sample/user-data and recommend the class label based on the features.
- Making the prediction of class labels using range set data so that the class labels can be predicted based on range query. For example: Predicting salary for people of age 35-40 and working for 20-30 hours-per-week.