Data Visualization

Communicating Predictive Analytics

James William Dunn
jwdunn2@asu.edu

I. Introduction

The following report details a collaborative team project to analyze a dataset, create statistical models of income prediction, and construct data visualizations to communicate the results. Given the scenario of a college seeking to boost its enrollment figures using the US census, we followed a process of examining four components: data, tasks, stakeholders, and visualizations as described by Fisher and Meyer [1]. We performed exploratory data analysis to identify significant attributes and their correlation to salary class. From the scenario parameters, we partitioned the work into six manageable task units connected to the college's business goals. We identified the stakeholders and audiences. We determined the inference rules around the dataset features and chose the types of graphics required to convey these rules.

II. Solution

We began our group experience by brainstorming a team identity and selecting *cubic engineers* to reflect six unique perspectives. We then analyzed the *customer ask* which states that a fictional UVW College board is challenging its marketing department to boost enrollment figures by using a census dataset to predict the factors that determine a person's salary level. This translates into a high-level business goal of their marketing team using a new software application to explore the demographics of prospective students. We were tasked with identifying those predictors and creating two reports: First, a succinct report aimed at the college executives who would make a decision about funding the development of the application for the marketing department. Second, to compile a detailed systems report directed at our coworkers who would be developing the application for the college.

Part of the customer request is to use the US census data from 1990. Technically, the dataset from UCI [2] is an extract of 48,842 records that reflect adjusted gross income of more than one dollar. Each row contains 9 categorical data fields and 6 continuous numeric fields.

We began our exploration of the census dataset distributions and correlations using an interactive pandas DataFrame profiling tool[1]. In addition, we explored the dataset using standard tools such as Python and Google Sheets, building out a number of preliminary graphs (see Fig. 1).

Figure 1: Exploratory data analysis sketches.

Also, we trained a machine learning algorithm to create a decision tree. From this we investigated the primary branching factors. For instance, if the prediction error is 15% and the main determinants are age and education, this may inform our search for concrete relationships among the features. We also concluded that if the development team were to build an application to predict income based on the factors we identify, then perhaps we can further assist by creating a low-fidelity prototype. This took the form of a rudimentary console application which poses fourteen questions and uses a decision tree to predict a salary class.

Another process we explored was the user-story concept. Initially, we understood this to be synonymous with a marketing profile, thus we created several prospective student personae as an attempt to provide insight into how the data connects to the client. Later, we gained a different meaning, that of the agile software development process: a user story is an informal explanation of a software application feature from an end-user's perspective[2].

During our investigation, we learned that the UCI data extract does not accurately reflect the full US population. For example, the 1990 census reports[3] that the population was 51% female whereas the extract reflects a composition of only 33% female. A case study [3] at MIT investigated that bias. One of the metrics that we noticed in the extract was that only 11% of the female population reported a salary over $50K. Based on this data, we can statistically infer to a larger population that 9 out of 10 females earn less than $50K. We found our first strong predictor of income level.

One question we posed was: how do we know how many visuals to create? 5, 10, 15, 20? We decided that six slides is sufficient to convey a compelling message to an executive board. Two minutes per visualization translates to a twelve minute presentation. Additional visuals act as supporting evidence. This meant finding six predictors. Also, we considered what makes a compelling majority – we agreed

1 github.com/pandas-profiling

2 atlassian.com/agile/project-management/user-stories
3 census.gov/library/publications/1993/dec/cqc-03.html

that a ratio of 4:1 (or 80%) is *good*. Above this would be *strong* and below this we consider *weak*.

We iterated on combinations of the fifteen fields of the dataset until we identified six majorities. For example, education alone is correlated with earnings potential. Also, when we plotted a box-and-whisker visualization of age and relationship categories for both salary classes, we noticed one value, in particular 'own-child', does not have an overlapping interquartile range – all the other values do. We concluded that, given a new record with this pair of features, one can determine the likely salary class.

III. Results

Decision tree factors

We achieved 86% accuracy in predicting salary class with our decision tree model. In descending order, we determined these branching factors: relationship, education-num, hours-per-week, age, and occupation.

The statistical predictors

1. When education level[4] is 10 and lower, there is a higher probability (2:1) of earning under $50K, while those at level 13 and above are more likely (3:1) to earn over $50K.
2. If the gender field is *female*, there is an 89% probability that earnings are less than $50K.
3. If the relationship field is *own-child* and age is in the range of 19 – 27, then there is a 56:1 salary class ratio for earnings less than $50K.
4. If the age field is 23 – 25 and education level is 10 or 11, there is a 10:1 salary class ratio in favor of earnings less than $50K. If the age range is 41 – 43 and education is 12, there is a 3.6:1 salary class ratio for over $50K.
5. When the hours-per-week field is less than 40, then there is a 3:1 salary class ratio for earnings less than $50K.
6. Where the native country field contains values other than US, Canada, Mexico, or Germany then there is a 2:1 ratio of individuals earning less than $50K.

The visualizations (see Fig. 2)

1. To compare education levels and earnings, we originally proposed a box-and-whisker chart and later concluded that the two salary classes were not separated enough. Instead, we selected a line graph to present a compelling contrast of values. We graphed education level on the x-axis and percentage of instances on the y-axis. A green line traces the over-50K and a blue line traces the under-50K.
2. To represent the 9:1 ratio of female earnings, we utilized a donut chart composed of blue and green. This clearly communicates the majority class that earns under $50K.
3. For the relationship visual, we utilized a box-and-whisker plot, a blue one for the under-50K and a green one for the over-50K. The interquartile ranges do not overlap, conveying a visual distinction between the classes. We oriented the boxes on the horizontal with age mapped along the x-axis.

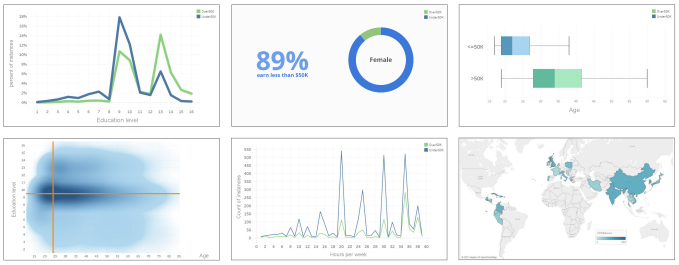

Figure 2: Selected final data visualizations

4. Next, a scatter plot of blue dots indicate the distribution of under-50K salaries with age on the x-axis and education level on the y-axis. The count of records map to oversized density markers with 65% intensity. We overlaid orange lines to identify the center of the resulting cluster.
5. To highlight the opportunity in the part-time work sector, we zoomed in on a line graph in the under 40 hours region. The hours-per-week run along the x-axis. The count of instances map to the y-axis. The data was extracted from a normalized salary class subset.
6. To visualize the counts of individuals from the native countries, we initially experimented with proportional symbol mapping, however, we decided a choropleth map felt more natural. A sequential color scheme [4, p. 6] indicates the number in each region, again using blue to echo the under $50K salary class.

IV. Contributions

On this project, I performed a dual role as both principal data analyst and project manager. Leveraging my professional experience as a relational database consultant, I focused on understanding the dataset and identifying the key factors for predicting income based on the *customer ask*. For an alternative perspective on the dataset, I investigated the decision tree that my teammate Lu Gao created using the census dataset. I wanted to know what the machine learning algorithm had determined to be the main branching factors during its transit toward a leaf node. With Google Colaboratory[5], I was able to print the tree node data and connect the encodings back to the original features. Using Jupyter, I created a mosaic plot to compare how the values in the 'relationship' feature correspond to income level.

In addition, as a seasoned project manager, I offered the team guidance, structure, and task definition. Considering the standard project vectors for success, I planned the scope of work and resources to meet the fixed schedule. As we moved forward, I tracked the team's progress on the various tasks such as exploratory data analysis, content contributions, and reviews of the reports. Serving as the project librarian, I collected team member findings, code, and corresponding visuals. Also, I set up the report outlines and refined the layouts as information mounted. I designed both reports to showcase the data visualizations with supporting evidence surrounding each piece. I followed the Harvard guidance [5] for slide shows to busy executives which suggests keeping the slide count to a minimum.

4 Education levels: 10 = some college, 11 = vocational associates, 12 = academic associates, 13 = Bachelors.

5 colab.research.google.com

In order to drive the priorities around the exploration of income factors, I suggested we focus on the user-story concept. Initially thinking that *the user* was a prospective student of the college, I created and shared example marketing profiles such as: *Bert is a retiree with a college degree and extra time on his hands*.

As we made further progress, it became clear that the marketing department would be using the application which XYZ Corporation may be developing and maintaining, depending on stakeholder approval after the executive briefing. Through careful analysis of the *customer ask* and all project description documents, I was able to sketch a flow chart of the process which I refined into a cycle diagram (see Fig. 3). This was included with the systems documentation report and helped to clarify the project roles and responsibilities for each of the four fictitious groups involved:

1. *stakeholders*: the UVW College Board
2. the *user*: the UVW marketing department
3. *cubic engineers*: the XYZ Corporation analysis team
4. our *coworkers*: the XYZ development team

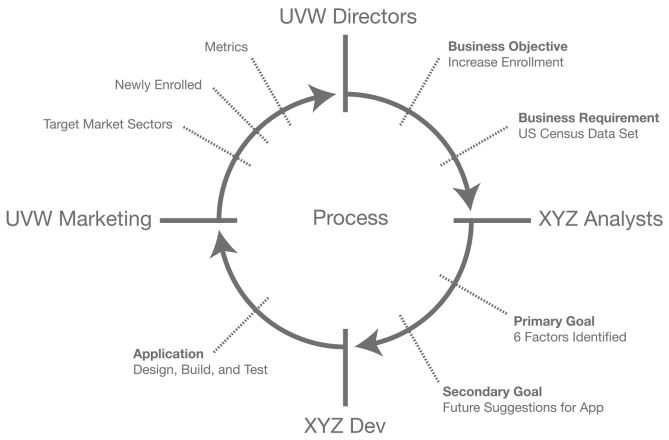

Figure 3: Cycle diagram designed by James Dunn

During my exploration of the data using tools such as Python, Google Sheets, Excel, and Jupyter Notebooks, I decided to leverage Tableau[6] in my search for correlations. This exercise led to seventy-four visualizations using Tableau, three with Jupyter, and another twenty with Excel.

As team members shared their findings, I verified the analysis logic, extracted the salient points, and rendered similar diagrams in Tableau for a cohesive appearance in the milestone reports. For example, I reconstructed the pairing of hours-per-week vs income and the combination of education vs income.

I value professional aesthetics. A team identity is not only the work accomplished but how it is written and presented. To ensure a cohesive presentation, I assembled a reporting framework (see Fig. 4) and directed the primary factors of visual communication—from layout and typography to a synchronized color treatment of the salary classes: blue for under-50K and green for over-50K.

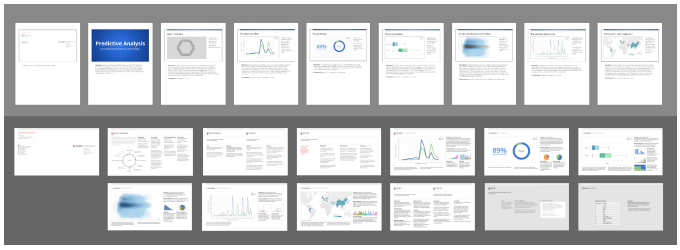

Figure 4: Executive slides and systems documentation layout

V. Lessons Learned

This project challenged me in two ways: 1) the generality of the project description, and 2) the Tableau learning curve. The first challenge was unpacking the complexity in the project briefing, similar to a real-world project. I found a need to synthesize a story around the whole process—to place all the characters on a virtual stage and to learn the role I needed to play as a director and analyst in connection with making a rapid delivery. I replaced UVW with ASU and replaced fictitious user stories with neighbors, friends, past co-workers, and family who fit into the persona of the factors identified. This grounded the requirements, and in the end, the predictor system works for and aligns to real-world individuals. With current census data, any university or college might consider leveraging this form of marketing to boost enrollment.

The second challenge was the need to rapidly learn a complex visual analytics system. Creating visualizations with Tableau proved to be more difficult and time-consuming than initially expected. The *Show Me* feature does not offer guidance other than a hint. I repeatedly explored the help feature and searched the Internet to learn new procedures. An improvement would be an option within Tableau to interactively step a user through a target chart type.

I see a direct application of what I've learned in this activity to a current project on which I am engaged as a volunteer at kynamatrix Research Network[7]. The Proximity Algorithm is a scalable system of assisting survivors of natural disaster by creating temporary, well-organized urban environments. Data analysis and visualization play a crucial role in managing these large structures through the use of dashboards to display metrics for understanding the operational status and trends.

Through this project and coursework, I have gained a deeper appreciation of the importance of analysis, statistical modeling, and data visualization. Creating an accurate visualization requires a solid understanding of its underlying dataset, its feature relationships, both obvious and discoverable, and a recognition of the potential need to normalize and/or scale the data. Equally important are the design decisions for presentation.

The six-member *cubic engineers* analysis team included: Hongwei Liu, Adrian Pantea, Getachew Ali, Marmeena Beniamin, Lu Gao, and James Dunn.

6 tableau.com

7 kynamatrix.org

VI. References

[1] D. Fisher and M. Meyer, *Making Data Visual.* O'Reilly Media, Inc., Sebastopol, California, 2018.

[2] C. Blake, C. Merz, UCI repository of machine learning databases [Online], Department of Information and Computer Science. University of California, Irvine, 1998. archive.ics.uci.edu /ml/datasets/adult

[3] R. Fletcher, D. Frey, M. Teodorescu, A. Gandhi, and A. Nakeshimana, *Exploring Fairness in Machine Learning for International Development*. Spring 2020. MIT bit.ly/3jnYgRz

[4] R. Maciejewski, *Data Representations, Transformations, and Statistics for Visual Reasoning.* Synthesis Lectures on Visualization, 2(1), pp.1-75, 2011. DOI: 10.2200/S00357ED1V01Y201105VIS002

[5] N. Duarte, "How to Present to Senior Executives," [Online], Harvard Business Review, 2012. hbr.org/2012/10/how-to-present-to-senior-execu