University Projects

A Highlight of data analytics/business analytics projects I have work in the past 1 year (top projects executed, July 2023 – Till present).

A. Predicating Loan Default for Mammoth Bank

Objective:

The primary goal of this project was to develop a predictive model to identify customers at Mammoth Bank who are likely to default on their loans. This enables the bank to extend more loans confidently while minimizing default rates.

Approach:

1. **Exploratory Data Analysis (EDA):**
   - Analyzed the `loan_default` dataset, which included 1000 observations and 15 predictor variables.
   - The response variable `Default` was binary (1 for default, 0 for non-default).
   - Identified and handled outliers in variables such as Checking_amount, Term, Credit_score, and others.
2. **Data Preparation:**
   - Split the data into training (80%) and testing (20%) sets.
   - Performed 5-fold cross-validation to ensure model robustness.
3. **Modeling Techniques:**
   - Implemented Subset Selection (Backward Elimination and Forward Selection) and LASSO Logistic Regression.
   - Backward Elimination and Forward Selection aimed to select the most significant predictors by minimizing the Akaike Information Criterion (AIC).
   - LASSO (Least Absolute Shrinkage and Selection Operator) was used to prevent overfitting by penalizing the absolute size of the regression coefficients, leading to a sparse model.
4. **Evaluation:**
   - Compared models using Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curves.
   - Forward Selection achieved the highest AUC (0.9827), followed closely by Backward Elimination (0.9826) and LASSO Logistic Regression (0.9824).

Results:

- The most impactful predictors for loan default included Age, Checking Amount, Saving Amount, Credit Score, Home Loan, Personal Loan, Term, Education Loan, and Amount.

- The final recommendation for Mammoth Bank was to use the Forward Selection model due to its slightly higher AUC, ensuring reliable prediction of loan default.

Tools and Algorithms:

- R programming language and the `glmnet` package for LASSO Logistic Regression.
- Subset Selection methods for feature selection.
- ROC curves for model evaluation.

Conclusion:

This project provided Mammoth Bank with a statistically validated model to predict loan defaults, allowing the bank to make informed lending decisions, reduce default rates, and increase profitability.

B. Project Title: Predicting Bordeaux Wine Prices and Quality

Objective: The aim of this project is to predict the quality and prices of Bordeaux wines using an econometric model based on weather conditions. The goal was to provide an accurate prediction method that bypasses the subjective opinions of wine experts, thereby offering a more objective and cost-effective approach to assessing wine value.

Methodology:

1. **Data Analysis and Preparation:**
   - Utilized a dataset containing 27 red Bordeaux vintages with variables such as year of harvest, logarithm of average market price, winter rainfall, average growing season temperature, harvest rainfall, wine age, and France's population during the harvest year.
2. **Multiple Regression Analysis:**
   - Performed multiple regression using R to analyze the relationship between wine prices and various weather-related factors.
   - Included quadratic terms in the regression equations to account for non-linear relationships.
   - Conducted an F-test to compare models and determine the best fit.
   - Corrected standard errors for heteroscedasticity to improve model reliability.
3. **Visualization:**
   - Created scatter plots and regression lines using ggplot2 to visually represent the relationships between variables.
   - Differentiated wines above and below average price using color-coded labels.
4. **Regression Models:**
   - Estimated multiple regression models, including robust standard error adjustments.
   - Generated regression tables using the `modelsummary` package.

- Calculated marginal effects and group means to provide deeper insights into the data.

Algorithm Used: Multiple Linear Regression, with quadratic terms and robust standard error adjustments.

Results:
The analysis showed a significant correlation between wine prices and weather conditions, specifically average growing season temperature and harvest rainfall. The best-fitting model, which included quadratic terms, explained 82.8% of the variability in wine prices. The results highlighted the potential impact of climate change on the wine industry and suggested that weather conditions could be used to predict wine quality and market prices effectively.

Key Skills Demonstrated:

- Proficiency in R for data analysis and visualization.
- Ability to perform multiple regression analysis and interpret complex models.
- Experience with econometric techniques and correcting for heteroscedasticity.
- Strong data presentation skills using ggplot2 and model summary for creating clear and informative visualizations and tables.

Tools and Technologies:

- R (ggplot2, lm, model summary, margins)
- Data visualization - Econometric modeling
- Statistical analysis

This project showcases my ability to apply advanced data science techniques to real-world problems. making it a valuable addition to my resume for a Data Scientist role.

C. Project Title: Analyzing Voter Turnout Using Social Pressure Interventions

Objective: The project aimed to investigate the effects of social pressure and civic duty reminders on voter turnout in the 2006 primary election in Michigan. The study sought to understand why people vote despite the seemingly irrational cost-benefit analysis of voting.

Methodology:

1. **Data Analysis and Preparation:**
    - Utilized a dataset from a field experiment involving nearly 350,000 voters who were randomly assigned to one of four treatment groups or a control group.
    - Loaded the replication data for Gerber, Green, and Larimer (2008) using R.

2. **Experiment Details:**
   Treatment Groups:
   - Civic Duty: Mailout reminding voters that voting is a civic duty.
   - Hawthorne: Mailout informing voters that researchers would study their turnout.
   - Self: Mailout displaying the household's prior election turnout record.
   - Neighbors: Mailout displaying the household's and neighbors' prior election turnout records.
   - Control Group: No mailout.
3. **Statistical Analysis:**
   - Calculated turnout rates for each group and the number of individuals allocated to each group.
   - Created new variables for gender, age, and the 2004 election turnout to conduct balance checks.
   - Performed regression analysis to estimate the average treatment effects of different interventions, both with and without controlling for demographic factors.
4. **Difference-in-Differences Analysis:**
   - Estimated treatment effects separately for men and women to observe potential differences in impact.
   - Conducted a placebo test using pre-reform data to check for any spurious effects.

Algorithm Used: Fixed Effects Model and Difference-in-Differences Analysis were employed to analyze the impact of the treatments.

Results:
The treatments significantly influenced voter turnout, with the "Neighbors" treatment showing the highest increase in turnout. Gender did not significantly moderate the treatment effects, indicating that interventions were equally effective across genders. The balance checks confirmed that the randomization process was successful, with no systematic differences between treatment and control groups. Difference-in-differences analysis showed that the treatments led to a statistically significant increase in voter turnout.

Key Skills Demonstrated:
- Proficiency in R for data manipulation, statistical analysis, and visualization.
- Experience with fixed effects models and difference-in-differences analysis.
- Ability to conduct and interpret balance checks to ensure the validity of randomization.
- Strong understanding of experimental design and causal inference in social science research.

Tools and Technologies:

- R (lm, model summary, ggplot2, fixist)
- Data visualization
- Statistical analysis
- Experimental design

This project demonstrates my ability to apply advanced statistical techniques to real-world social science problems, making it a valuable addition to my resume for a Data Scientist role.

D. Analyzing the impact of Abduction on Education, Distress, and wages in Uganda

Objective:

The project aimed to estimate the impact of forced military service by the Lord's Resistance Army (LRA) on various outcomes, including education, emotional distress, and wages, among male youth in war-affected regions of Uganda.

Methodology:

1. **Data Analysis and Preparation:**
   - Utilized a dataset from a panel survey of male youth in Uganda, focusing on those abducted by the LRA.
   - Variables included abduction status, years of education, emotional distress index, log of average daily wage, and several demographic indicators.
2. **Naive Average Treatment Effect (ATE) Calculation:**
   - Ran separate regressions to estimate the naive ATE of abduction on education, distress, and log wages.
3. **Propensity Score Matching:**
   Used a logistic regression model to calculate propensity scores for each individual to be abducted, based on covariates such as age, parental education, household size, and location indicators.
   Applied optimal matching to estimate the ATE using propensity score matching for the three dependent variables.
4. **Balance Check:** Conducted a balance check using a Love plot to ensure covariates were balanced between treated and control groups after matching.

Algorithm Used:

Linear Regression for naive ATE estimation.

Logistic Regression for calculating propensity scores, and

Propensity Score Matching for ATE estimation.

Results:
Naive ATE Results: Abduction had a significant negative effect on education (-0.595 years) and a significant positive effect on distress (0.593 points).
The effect on log wages was not significant.

Propensity Score Matching Results:

- After matching, abduction was still associated with a significant decrease in education (-0.484 years) and a significant increase in distress (0.790 points).
- The effect on log wages remained insignificant.

Key Skills Demonstrated:

- Proficiency in R for data manipulation, statistical analysis, and visualization.
- Experience with logistic regression and propensity score matching.
- Ability to perform and interpret balance checks using Love plots.
- Strong understanding of causal inference methods in econometrics.

Tools and Technologies:

- R (lm, glm, MatchIt, cobalt, modelsummary)
- Data visualization
- Statistical analysis
- Causal inference techniques

This project highlights my ability to apply advanced econometric techniques to analyze the impact of significant social issues, making it a valuable addition to my resume for a Data Scientist role.