Hayden Allison

2/4/2024

Proposal

1. Project Overview
   1. Are classifications or measurements evaluated by Prosper regarding loan applicants utilized in distributing them corresponding conditions? *This question wasn’t very successful in terms of providing Prosper business value. Therefore, an extension is added that supplies value, which is:* *can loan outcomes be predicted for new customers who apply?*
      1. Prosper has a wide demographic of customers, in terms of: age, income range, occupation, credit score, etc. The loans approved have a massive impact on Prosper’s profit margins. These customer measurements should be evaluated against the company’s internal, historical data for distributing a proven-to-work set of loan conditions. For example, Prosper definitely does not want to be handing low APR’s and high loan amounts to individuals with high credit balance and delinquency amount.
      2. New customers that apply will result in: defaulted, charged off, current, cancelled, completed, past due \_\_ days. When a model assists in the decision-making process for approving customers, it provides cost optimization by suggesting denial of higher risk customers (defaulted and charged off) based on Prosper’s records of completed loans. The current and cancelled loan records are not analyzed in this portion of investigation, as they are incomplete.
   2. Loan approval strategy success is awarded based on foresight and caution. These are only accomplished when an examination of both the customer and previous outcomes of customers with similar demographics is conducted. Any financial organization or bank understands that loans can be a complete miss. Situations happen where the loan is defaulted on, which can result in a total loss for the bank. Quantifying the measurements of instances like these, also of successful instances where the loan is paid in full, can create a model to predict the outcome or assess the risk of a specific customer. When this approach is taken, unrealized profit can be obtained while risk is mitigated. Industries around the globe classify customers in this manner. Prosper should have a similar process given its eligibility to do so. This project evaluates Prosper’s current model and provides a model that predicts if a customer’s loan will be a success or failure.
   3. Summarization of 3 works
      1. *Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm*
         * SUMMARIZATION: Loan approval conditions are examined such as employment type, gender, and more.[3] These customer demographics along with aggregation of loan outcomes are used for building a Random Forest model to predict suitable customers for loan approval. It also has an end result of applicant’s eligibility and risk mitigation to the bank. The Random Forest model selection was based on its higher score.[3] Scores were pulled from a previous study comparing decision tree, SVM, AdaBoost, Bagging, and Random Forest to a Logistic Regression benchmark.[3]
         * DEVELOPMENT IMPACT: It’s a useful reference for the evaluation of a random forest or decision tree model. The loan outcomes model I’m going to deploy may best be suited for a Decision Tree or Random Forest. The findings in this study are factored into my model planning phase given the in-depth information of these two specific candidate models.
      2. *Prediction of Loan Approval in Banks using Machine Learning Approach*
         * SUMMARIZATION: A research study which explores machine learning models for classifying loan applications for approval.[2] These included KNN, Decision Tree, Random Forest, and Naïve Bayes. They were individually scored primarily based on accuracy, with Naïve Bayes achieving 83.73%.[2] Python, R and Jupyter Notebooks were mentioned as tools. Instructions are included that further describe how to preprocess, in-process, and postprocess. Recommendations are given for assisting readers in developing their suitable model.[2]
         * DEVELOPMENT IMPACT: My analysis was conducted using Python and Jupyter Notebooks like this study. The same machine learning library (sklearn) was even used. This is invaluable for depicting a proven-to-work representation of their methods, one I’m able to closely replicate. Using this research study as a reference, I can train an outcome model for each of the 4 model types included within. I have these analysts’ results as a baseline when testing.
      3. *Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis*
         * SUMMARIZATION: Machine Learning models are initialized for predicting the eligibility of loan applicants based on historical data from previously approved applicants.[1] A comprehensive set of models are used, including RF, LR, XGBoost, Adaboost, Lightgbm, DT, and KNN.[1] LR scored the highest with an accuracy of 92%.[1] The literature review conducted has scores listed for model accuracies of previous analyses. Postprocessing evaluations are charted with ROC charts, AUC analysis, and accuracy scores. A multitude of different measures are applied for determining effectiveness: accuracy, sensitivity, specificity, precision, F1-score, and AUC.[1]
         * DEVELOPMENT IMPACT: No mentions of tools involved within my project are included in this research study. However, it does mention new models that haven’t been mentioned by the other research studies. LightGBM, AdaBoost, and XGBoost can be considered as candidate models as well. Substantiating evidence for a model’s effectiveness is done thoroughly within this particular research study, and those same measurements can be applied for my loan outcome models.
   4. Deliverables
      1. My narrative of key findings and quantified business value is to be presented in a PowerPoint presentation, a static report, and Exploratory Analysis. Visualizations and concise statements are chosen methods of communicating results. Two models will be deployable: one for imputing customer’s missing Income Range and Employment Status, then an additional model for predicting the outcome of a customer’s loan. The 1st model is for occurrences of record incompleteness. The 2nd model assists in the approval process – if a customer is classified as probable to default on their loan, then their application should be denied and vice versa.
   5. Analytic solution benefit
      1. The solution to my analysis is beneficial to the organization in terms of risk mitigation, cost optimization, and customer satisfaction. Risk mitigation is procured through predicting the outcomes of high-risk customers and denying their application. Cost optimization applies from predicting outcomes of low-risk customers and accepting their application. Customer satisfaction is amplified through means of assigning them loan conditions tailored to their qualities. Under this technique of assigning loan conditions, the average and above customers receive favorable conditions and are more likely to consider Prosper for future loans. A crucial decision is commonly made in financial institutions that lend money. Is a customer worthy of approval, or should they be rejected? Failure to make the proper choice restricts profit immensely. The loan outcome model has the ability to accurately predict. Stakeholders can deploy this model for assisting this decision-making process to: capitalize on loaners that are likely to complete their loan and avoid loaners that are likely to miss payments or default.
2. Project Plan
   1. Goals, Objectives, Deliverables
      1. GOALS:
         1. Develop a model that accurately predicts customers that are likely to complete their loan and those likely to default or become charged off.
         2. For the crucial customer qualities of Employment Status and Income Range, initialize a model that imputes the correct value.
         3. Find a relationship between customer qualities and their corresponding loan conditions.
         4. Determine if Prosper’s classification system of risk (Prosper Score) is affected by the customer demographics.
      2. OBJECTIVES:
         * GOAL #1: Develop model suitable for deployment and acts as a solution.
           + Train compatible classification type candidate models.
           + Test the selected candidate models.
           + Measure the effectiveness of each model.

Accuracy, sensitivity, precision, F1-score, and ROC charts.

* + - * + Optimize suitable model by removing unnecessary attributes.
      * GOAL #2: Develop model for imputing missing values for Income Range and Employment Status.
        + Identify a model compatible with a bi-feature dependent variable input.
        + Train and test the model.
        + Measure the effectiveness with accuracy.
        + Optimize model by removing unnecessary attributes.
        + Impute the values classified by the model in records containing NULLs.
      * GOAL #3: Communicate key findings and business value through visualizations.
        + Visualize distributions or frequency of variables with a variety of chart types.
        + Aggregate data based on the quality showing distribution of the loan condition variables.
        + Compare categories of a specific customer quality in terms of frequency and other measurements.
        + Run ANOVA tests to compare the means of loan conditions (APR and amount) per quality value (Income Range and Credit Score). Assess statistical significance with P-value. A total of 4 ANOVA tests must be completed.
      * GOAL #4: Formulate an answer for the relationships between Prosper Score and other features.
        + Visualize customer demographic distributions for each Prosper Score.
        + Compare the distributions and frequencies between Prosper Scores for various customer demographics.
      * Clean & assess data
      * Provide recommendations for Prosper from key findings.
      * Quantify business value derived from predictions of applying customers as likely to complete or likely to default.
    1. DELIVERABLES:
       - PowerPoint presentation
       - Static report
       - Explanatory analysis
       - 2 models ready for deployment
         * Impute model
         * Prediction of loan outcomes model
  1. Scope
     1. Scope is described in various ways throughout this project plan. A majority of the components are already covered. Constraints, boundaries, and dependencies.
        + CONSTRAINTS: no available GUI application for data analysis. All visualizations will be created via Python code as I’m more proficient in Python than R. Records don’t have a date associated with them.
        + BOUNDARIES: My analysis is focused on loan conditions & outcomes, customer demographics and Prosper Score. Not all attributes of the supplied data set are explored due to their irrelevance. Improving customer satisfaction and overall experiences is the first area of focus. Risk mitigation and cost optimization for Prosper is the second.
        + DEPENDENCIES: Deliverables require the discovery of key findings and relationships. Quantifying the business value relies on numeric measurements from the results of loan outcomes model.
  2. Methodology - CRISP-DM
     1. DISCOVERY: Formulate a research question. Research of published projects or studies related to my research question are explored – the 3 in my reference section. These provide techniques used by other analysts, along with instructions. I gain an understanding of the benefits this analysis will have for Prosper (Part 2) and the customer (Part 1). Then, with this knowledge, I create an analytics framework contrived of a hypothesis and areas of focus for finding results. Selection of tools, environment, and equipment take place.
     2. DATA UNDERSTANDING: I begin by creating a subset of the original Udacity dataset with variables that are relevant to my areas of focus. Through data profiling, I’m able to familiarize myself with the dataset. Referencing the documentation, relevant variables can be exported into my analysis subset. Definitions are read through for transparency of the function regarding each variable.
     3. DATA PREPARATION: Cleaning and assessing the data. Data veracity is gauged through completeness, accuracy, relevancy, validity, and uniqueness. I check for values that change may over time such as Prosper Score (if its dependent on an updating credit score for example), outliers, skewed distributions, missing values, and improper encodings.
     4. MODEL PREPARATION: An exploratory analysis is now conducted to discover relationships, patterns, determine if the research question can be answered, and substantiate evidence for picking suitable candidate models.
     5. MODEL EXECUTION: Training, evaluating, and testing of the candidate models occur. I optimize datasets of the model to increase effectiveness. From measuring the models’ effectiveness, I choose the 2 models that are most accurate. Determine if they require a different environment, additional/less inputs, or further evaluation to circumvent overfitting.
     6. COMMUNICATE RESULTS: Select key findings, quantify business value, and provide results in various report types (presentation and paginated report). Determine if the project was successful in finding a solution using the measurements of success.
  3. Timelines and Milestones
     1. Sourced from a previously completed Udacity project. These are the milestones left since a sizable portion of work has already been completed.
     2. 2/4/2024: Research published projects and studies. Establish measurements of success. Duration of 1 day.
     3. 2/5/2024 – 2/8/2024: Finish exploratory data analysis and identify compatible models. Duration of 3 days.
     4. 2/9/2024 – 2/10/2024: Train, evaluate, and test candidate models. Optimize the most effective model for the 2 different model goals. Duration of 2 days.
     5. 2/11/2024 – 2/14/2024: Summarize findings in a PowerPoint presentation, Explanatory Analysis, and Tableau Dashboard. Duration of 3 days.
  4. Resources & Costs
     1. Personal desktop computer: $2000
     2. Python & Jupyter Notebooks: Free
     3. Internet search engine: Free
     4. Microsoft Office: $7 monthly
     5. Scientific calculator: $10
     6. ANOVA F-Statistic table: Free
     7. Work hours: 72 hours @ $20 an hour - $1,440
  5. Criteria for Success
     1. Are the differences of means regarding Income Range and Credit Score statistically significant among their respective categorical values for customer loan amounts and APR?
        + The appropriate statistical model (ANOVA) was chosen for evaluating this.
        + The 2 most significant customer qualities were selected - the likelihood of having an impact on dependent variables.
     2. The imputation model provides a higher success probability of classifying missing values than random sampling.
        + Imputes for both Income Range and Employment Status.
        + Decide the correct value from a range of class members.
     3. The outcome model provides a higher success probability of predicting customer’s loan outcomes than random sampling.
        + Predicts the highly influential profit outcomes: completed, defaulted, and charge off.
        + Assists stakeholders in the decision-making process regarding loan approvals from the ability to correctly predict.
     4. The final outcome model has effectiveness measurements similar to models referenced in the research studies.
        + F1-score, accuracy, recall, and precision.

1. Solution Design
   1. Hypothesis
      1. Null: Uniform average/median for the entire class.
      2. Alternate: The average/median of loan conditions (APR and amount) is relative to class members.
   2. Method
      1. Descriptive & Predictive
      2. I’ll be examining historical data of Prosper’s customer loans to identify relationships and patterns. Part of the analysis uses this information to understand what happened in terms of loan conditions for certain customer qualities. Another part uses it to evaluate the risk assessment effectiveness the Prosper Score provides. These first two parts don’t investigate causation behind distributions and measurements. They’re piecing together a detailed picture that doesn’t go further than a traditional descriptive analysis. However, from dissatisfaction with the results and a requirement to provide quantifiable business value, the final part is added. It uses the same historical data to predict the loan outcomes of future customers for amplifying profits. This part is inherently predictive because of the model’s ability to suggest an action based upon an unknown variable (the loan outcome).
   3. Tools & Environments
      1. Python (coding environment)
         * Libraries: scikit learn (machine learning models), numpy (numerical computing), pandas (data manipulation), seaborn (visualizations), matplotlib (visualizations)
      2. Jupyter Notebooks (coding environment)
      3. Microsoft Word (documentation) & PowerPoint (presentation)
      4. Microsoft Windows 11 (operating system)
      5. Google Search Engine (research tool)
   4. Output methods and metrics
      1. Model accuracy substantiates a proportion for correctly classified instances out of the total number of instances in the dataset. Represented as a number from 0 – 1.
         * Accuracy is a justified metric since it provides an easy-to-understand metric for the overall effectiveness of a model. Having the prediction contrasted against the actual classifications of the training set is the main metric for identifying a suitable model.
      2. Model recall measures the ability of the model to correctly identify all positive instances out of the total actual positive instances. It’s a narrower metric when compared to accuracy since a strict positive proportion is the output. Represented as a number from 0 – 1.
         * Recall is a justified metric since it quantifies the true positive classifications performed by the model. The relationship to Prosper’s revenue is found in the fact that not identifying customers likely to have a loan outcome of “complete” or “defaulted” or “charge off” harms overall profits
      3. Model precision identifies the number of correct positive instances out of all positive instances made from the algorithm. This is a more direct measure of model capability compared to recall as it’s solely focused on model instances instead of including dataset instances. Represented as a number from 0 – 1.
         * Precision is a justified metric for further in-depth information for positive classification. When dataset instances are replaced by model instances, the quality of predictions is now measured. It’s a different metric versus the correctness of predictions seen in recall or accuracy. Gauging the quality of predictions would be impossible without this metric.
      4. Model F1-Score is the harmonic mean of precision and recall. It’s represented as a number from 0 – 1.
         * F1-Score is justified because of its robustness to unevenly distributed data. In the likely event that there are classes or members with less frequency in the overall population, F1-Score represents a nuanced approach to depicting accuracy since it factors in imbalance/weighted inputs.
      5. Comparison of various models will take the metrics of F1-Score, precision, recall, specificity, and accuracy to identify which has the highest scores. Higher scores translate into a better model.
         * Metric comparison is a justified technique because it gives me the opportunity to train then evaluate multiple models instead of a singular algorithm. Each algorithm comes with strengths and weaknesses. Having a chance to weigh these characteristics means my solution will supply the most effective model. For Prosper, this has a direct impact on profit margins.
      6. ANOVA test is ideal for comparing the average of the dependent variables (loan amount and APR) against the members for the given classes of Income Range and Credit Score.
         * ANOVA is justified due to how it minimizes the number of tests required compared to t-tests or z-tests counterparts. Running a test for each member instead of the entire class would be utterly exhausting. The average comparison is vital in the evaluation of whether the 2 most significant customer qualities influence loan conditions under Prosper’s existing processes.
   5. Practical Significance
      1. ANOVA enables comparison of the f-statistic output to a critical value or using a p-value for measuring statistical significance. If it is shown through the f-statistic being a higher value than the critical value or the p-value is smaller or equal to my desired significance level, then that is evidence of the differences between averages is not to chance. What this comparison shows instead, is the dependent variables (loan amount and APR) are influenced by the independent variables (Income Range and Credit Score). For example, I can then confirm someone of a lower Income Range (such as $25,000-$50,000) has access to lower loan amounts due to their Income Range status, if their average for this range is lower than the higher ranges like $100,000+.
      2. The criteria for assessing if the expected benefit is provided for validating practical significance, is comparing the averages across independent variables category. For both Credit Score and Income Range, the lowest groups ($0 Income Range) should have the lowest average. It should simulate a hierarchy, where the higher up category values should have a higher average, until the highest category value is reached ($100,000+ Income Range) which should have the highest average. Note, this hierarchy is in ascend order for Loan Amount but descend order for APR since higher APRs should be reserved for riskier/less qualified customers.
      3. Decision-making is supported whether or not my null hypothesis is rejected. In the situation of such rejection, and the means of the category values function as a hierarchy, then that’s evidence of Proser assigning loan conditions based on the two significant customer qualities. Methods to refine this process can be recommended based on the averages provided by the testing. For example, if poor Credit Scores have lower APRs than fair, a recommendation of increasing the average APR of the poor value customers is able to be made. Prosper is then able to subsequently adjust their APR range for poor Credit Scores and fair Credit Scores. Also, there’s two outcomes which indicate Prosper isn’t properly assigning loan conditions. Outcome 1: null hypothesis fails to reject. Outcome 2: null hypothesis is rejected, but the hierarchy averages is not shown. In both of these outcomes, Prosper is advised to establish this hierarchy of loan condition values. With this information to guide them, employees at Prosper are to go through Income Range and Credit Score categories; then instantiate a range for loan amount and APR that corresponds with the category value. These ranges are advised to adhere to the same hierarchy mentioned previously. When this order is structured, Prosper can now assign a more lucrative value for loan amount and APR since it will mitigate risk of less qualified customers and bring in extra revenue from qualified customers.
   6. Visual Communication
      1. Customer APR by Loan Amount per Income Range – Scatter Plot:
         * Evidence for existence of a hierarchy
         * APR will be vertical axis and Loan Amount on horizontal axis
         * Color coded Income Range of all values for a 3D plot with a legend
         * Multivariate chart of population
      2. Average Customer APR by Income Range – Bar Plot:
         * Communicates averages for identifying hierarchy differences
         * APR on vertical axis and Income Range on horizontal axis
         * 6 columns in total: one for each income bracket
         * Includes margin of error and the alpha
         * P-value or F-statistic & critical value are within nearby vicinity for associating this test with statistical significance outcome
         * Univariate chart of population
      3. Average Customer Loan Amount by Income Range – Bar Plot:
         * Communicates averages for identifying hierarchy differences
         * Loan Amount on vertical axis and Income Range on horizontal axis
         * 6 columns in total: one for each income bracket
         * Includes margin of error and the alpha
         * P-value or F-statistic & critical value are within nearby vicinity for associating this test with statistical significance outcome
         * Univariate chart of population
      4. Customer APR by Loan Amount per Credit Score – Scatter Plot:
         * Evidence for existence of a hierarchy
         * APR will be vertical axis and Loan Amount on horizontal axis
         * Color coded Credit Score values for a 3D plot with a legend
         * Multivariate chart of population
      5. Average Customer APR by Credit Score – Bar Plot:
         * Communicates averages for identifying hierarchy differences
         * APR on vertical axis and Credit Score on horizontal axis
         * 5 columns in total: one for each credit score bracket
         * Includes margin of error and the alpha
         * P-value or F-statistic & critical value are within nearby vicinity for associating this test with statistical significance outcome
         * Univariate chart of population
      6. Average Customer Loan Amount by Credit Score – Bar Plot:
         * Communicates averages for identifying hierarchy differences
         * Loan Amount on vertical axis and Credit Score on horizontal axis
         * 5 columns in total: one for each credit score bracket
         * Includes margin of error and the alpha
         * P-value or F-statistic & critical value are within nearby vicinity for associating this test with statistical significance outcome
         * Univariate chart of population
      7. Model Effectiveness Measurements for Predicting Loan Outcomes – Table:
         * Contains the F1-score, accuracy, precision, and recall metrics from the chosen candidate model
           + A table for the worst performing candidate model is also displayed for depicting the effectiveness difference
      8. Loan Amount per Loan Status – Pie & Box Plot:
         * Quantifies the business value of outcome model. This value is depicted in terms of wrongful approvals (pie) and the revenue loss from these occurrences (box).
         * Univariate charts of a sample consisting of the following Loan Statuses: completed, defaulted, charge off
         * Pie chart is a breakdown of counts for each loan status type
         * Box plot shows loan amount on vertical axis and loan status types on horizontal axis
           + Visible line for median in each box distribution
      9. PowerPoint Presentation:
         * Slides will provide the framework for the above visualizations
         * Narration uses slides during the entire duration of the presentation
         * Crucial figures, key findings, and visualizations are given respective space
         * This serves as an audience optimized version of the final report
2. Datasets
   1. A previous class I completed for my degree with WGU through Udacity. It’s a publicly available dataset on Kaggle.com.
   2. To evaluate loan condition distribution based on customer classification, historical data from internal data systems under ownership of Prosper Loans is required. With this data, I can begin to investigate relationships and correlations between the two variables listed above. For the second part, my model uses finished loans to predict the outcome of future customers. If I’m unable to input previous successful and unsuccessful loans, my predictor model won’t be able train. For goals, the 2 types of models I’ll create for deployment have training and testing data. The third goal has customer demographics available for visualizing distributions and frequencies. The fourth goal has sufficient data available for modelling 4 ANOVA tests in detecting significant differences between the means of my variables. This dataset has decent data veracity for precision in my work.
   3. I downloaded 2 excel files through Udacity. One was documentation for the attribute definitions and the second was a table containing the ~111,000 records. The attributes contributing noise were removed and a third, final dataset was introduced that catered to the scope of my analysis.
   4. There weren’t veracity issues that were uncorrectable, outside of dates being unreported. Thankfully, Loan Status can be used to determine the current loans thanks to the “current” value – others signify loan completion. Around 100,000 NULL values existed in the dataset at first. Feature averages and modes were imputed to account for these. Regarding the features of Income Range and Employment Status, these had inputs from the Bi-Feature Multi-Class Logistic Regression model due to their perceived significance in my investigation. The model rectified the low chance mode or average had for providing the true value. A few attributes had outliers that weren’t far out of range while a couple were astronomically out of range. These couple were replaced with the mode/average. Four records had a “0” value for monthly loan payment with a status of “Final Payment in Progress”. Other records had a payment listed, so the average was imputed for these. Secondly with the payments, the prevalence of records containing an amount that didn’t make sense considering the Term and Loan Amount was high. Nearly all cases I viewed of this had payment that would never pay back the loan, given their term length. Monthly Loan Payment was removed due to this widespread inaccuracy.
   5. No personally identifiable information (PII) is contained in my dataset. Obvious sensitive financial information that is heavily regulated like credit card numbers, billing addresses, etc. aren’t present. These features were probably stripped before becoming a public dataset as mandatory per regulations that guide Prosper’s processes. There was an opportunity for financial sensitivity regarding the credit features – that had figures pulled from credit reports run on individuals. Laws demand not disclosing this information to the public under financial reporting restrictions. The privacy of individuals must be maintained in any financial sector. This requires keeping the original data secure and out-of-reach to the public. The code of ethics dictates that responsibility in this situation is adherence to the laws governing distribution of financial data, maintaining the privacy of individuals, and to provide an analytics solution that can fit within Prosper’s processes given the regulation imposed upon them.
      1. As a safeguard for legal requirements to not disclose the financial information of individuals, the credit features won’t be shown together in visualizations or tables and remains aggregated at all points for the audience. This assists in the ethics and privacy behind my project, as the individual records attached to a real person (even without the PII) are not identifiable by the audience. I’m able to communicate my findings to Prosper stakeholders or any public figure without intruding on the people’s privacy. Access to the original dataset is not permitted.
3. References
   1. [1] Al Mamun, M., Farjana, A., & Mamun, M. (2022, June). Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis. In *7th North American International Conference on Industrial Engineering and Operations Management*, <https://doi.org/10.46254/NA07.20220328>.
   2. [2] V, V., A.C, R., K N, V., & G, A. (2023, August 16). Prediction of loan approval in Banks using machine learning approach. SSRN. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4532468>
   3. [3] Dansana, Prasad, Mishra, & Patro. (2023, June). Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm. ResearchGate. Retrieved February 6, 2024, from https://www.researchgate.net/publication/371575364\_Analyzing\_the\_impact\_of\_loan\_features\_on\_bank\_loan\_prediction\_using\_Random\_Forest\_algorithm
   4. Data Analyst Online Course | Data Analytics. (2023, January 23). Udacity. <https://www.udacity.com/course/data-analyst-nanodegree--nd002?utm_source=gsem_brand&utm_medium=ads_r&utm_campaign=12712700850_c_individuals&utm_term=124530975510&utm_keyword=data%20analysis%20udacity_e&gclid=CjwKCAjw_aemBhBLEiwAT98FMk6-J8Ers7MLyK0Nu21U8NwiFUfa0_6XaPQrLLzNXOQxb41K_nW6UxoCxsUQAvD_BwE>