Hayden Allison

2/11/2024

Final Report

1. Project Overview
   1. Are classifications or measurements evaluated by Prosper regarding loan applicants utilized in distributing them corresponding conditions? Can loan outcomes be predicted for new customers who apply?
   2. Scope
      * The result of this journey can be reflected in 4 parts. Parts 1 & 2 are models suitable for deployment: model 1 is for predicting the outcome of new loan applicants, while Model 2 is filling in not available values within the dataset. Part 3 identifies whether Prosper Scores have noticeable relationships with other demographics. Part 4 is about quantifying analysis results to communicate business value regarding my proposed solutions. Prosper’s profit margins and their customers’ satisfaction are assumed to benefit from these.

Total price for resources and work hours is estimated to be $3,417. Cost effective tools and environments I relied on was Jupyter Notebooks with a Python Kernal, scientific calculator, Microsoft Office, Scikit Learn the machine learning library, Scipy the statistic model library, data-focused libraries of: Pandas, Numpy, Seaborn, Matplotlib, and lastly Google Search Engine to reference published literature and manuals.

* + - TIMELINE: These are the milestones left since work has already been completed.
      1. 2/4/2024: Research published projects and studies. Establish measurements of success. Duration of 1 day.
      2. 2/5/2024 – 2/8/2024: Finish exploratory data analysis and identify compatible models. Duration of 3 days.
      3. 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.
      4. 2/11/2024 – 2/14/2024: Summarize findings in a PowerPoint presentation, Explanatory Analysis, and a final report. Duration of 3 days.
    - CONSTRAINTS: No available GUI application for data analysis. Visualizations from code are created via Python. Table 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.
  1. Project Methodology: CRISP-DM
     + DISCOVERY: I start by formulating research questions. Published projects or studies related to the research questions are explored – which are 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 (Wijaya, 2022).
     + DATA UNDERSTANDING: Through data profiling, I’m able to familiarize myself with the dataset. Referencing the documentation, relevant variables can be exported into my analysis subset. Data definitions are read instead of making assumptions.
     + DATA PREPARATION: Cleaning and assessing the 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.
     + 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 (Wijaya, 2022).
     + 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 (Wijaya, 2022).
     + COMMUNICATE RESULTS: Select key findings, quantify business value, and provide results in various report types (presentation and final report). Determine if the project was successful in finding a solution using the measurements of success (Wijaya, 2022).

1. Project Execution

Overall, the execution felt relatively similar to the plan throughout the analysis. With knowledge of the dataset beforehand, my expectations were supported from those previous experiences. There wasn’t a moment where there was a lack of direction or focus. I knew what 4 goals that needed to be achieved and how to progress until a grand solution was found.

These goals remained unchanged throughout the project duration. I consistently worked towards uncovering the dataset’s intricacies to convert historical data into profit. Along the way, a couple tweaks were made to the objectives. Model metrics don’t include ROC charts, but instead have prediction results to rank against the true results. This was found to be more effective in evaluating effectiveness given the combination of used metrics. For the statistic tests done on my 2 independent variables, Credit Score and Income Range, and the 2 dependent variables, Loan Amount and APR, the pre-selected ANOVA test wasn’t compatible - other than a single test. It was exchanged for Kruskal-Wallis since the groups in 3 tests weren’t homoscedastic or approximately Gaussian distribution (proven with histograms & box plots). The criterion for success is seen per each corresponding directive and proved to be invaluable in fine-tuning success for the project.

CRISP-DM phases were almost fully adhered to. I omitted specific details about the steps that contrive every phase. Focus was placed on the general objectives which provided leg room to fill the steps in. Minor deviation took place in Model Execution phase, however. A proclamation of 2 suitable models were to be provided as components of a grand solution. Two of the predicting loan outcomes were extremely close in effectiveness, with their difference being attributed to the records’ characteristics within the input samples. Without further data to replace columns with ones that boost the success rate of one, both are equally recommended – for a total of 3 models.

The forecasted timeline is still holding true. It’s currently 2/11 and I’ll be creating this final report and a subsequent presentation through the next few days, with an approximate end date of 2/14. A singular adjustment was made to this flow - the remittance of a Tableau dashboard from time constraints. I don’t possess enough experience with the environment to develop a dashboard before training with online courses for up to 2 weeks. Extra emphasis is placed on a fully-fledged presentation, explanatory analysis, and final report.

1. Methodology

My plan was simple. Retrieve the downloaded files and select a range of variables to test and taper down to key variables. Ended up with a smaller range than I thought I would when first inspecting the dataset. Outside my range existed attributes that wouldn’t benefit this analysis in any way – like Estimated Yield and Group Key. Ideal for Prosper to have but useless for my purposes. I had hoped there would be more attributes allotted for customer qualities like age or gender. That’s the pronounced difference between my dataset and the studies I researched. They primarily focused on the customer while mine focused on credit factors.

An obstacle would be the variation I mentioned between mine and other analyses. I’m theorizing this didn’t complement the overfitting well. The influx of continuous quantitative attributes I was destined to analyze resulted in some models overfitting substantially. For this reason, I collected as many categorical variables as possible to reduce the number of continuous attributes.

Data governance had no hiccups. Customers are kept safe through total animosity. Visualizations contain aggregated data to further displace individual records away from themselves. The loan outcome model primarily uses information from credit reports and demographics (Income Range & Employment Status) to predict, which is permitted under Prosper’s regulations.

* 1. A disadvantage was the records not containing dates. Forming time periods for the dataset isn’t an available strategy. The single indicator that exists for separating current and completed loans was the value a record had in Loan Status. Ideally when configuring a predictor model, time periods are used for training and evaluation on distant past and recent past history. Then for predicting outcomes, the standard model uses recent past records to predict current or future events.

An advantage came in the form of a primary key saved in the dataset as Loan Number. I erased duplicated records based on this attribute, and it saved my analysis from being afflicted with repeat record syndrome. Without the advantage of having this attribute available, exactly 871 duplicate loans would’ve been included.

1. Data Preparation & Extraction

For data preparation, no major transformations or wrangling needed to be attempted. A few minor transformations for the loan outcomes model dataset did take place by using functions in Python’s Pandas library. I viewed modifications in Jupyter Notebooks to verify each.

Categorical attributes were numerically encoded exclusively using the Scikit Learn library. The first was Loan Status, which was encoded via LabelEncoder function given its status as the nominal, dependent variable. OneHotEncoder would’ve degraded the accuracy of every model. The next was Employment Status, which received OneHotEncoder treatment since it’s a nominal, independent variable. Lastly were Income Range and Credit Score. These underwent OrdinalEncoder for their ordinal, independent variable status.

The rest of attributes were numerical and standardized into a range of values relative to the individual attribute; using RobustScaler for its robustness against outliers from Scikit Learn. There was a handful of these with wide disparity. Features removed:

* Borrower APR
* Total Inquiries
* Bankcard Utilization
* Total Trades
* Debt To Income Ratio
* Prosper Score
* Total Credit Lines past 7 years
* Amount Delinquent
* Current Delinquencies
* Listing Category

These features posed too much risk of contributing to the overfitting problem - given their high standard deviations. Or weren’t contributing to the underlying pattern for successful predictions. These transformations helped prevent complex models (Random Forest, XGBoost, and Logistic Regression) from memorizing the noise to overperform in the training & evaluation sets, while enabling the other models (Decision Tree, ADABoost, and KNN) to predict more accurately (Ml/ai, 2023).

Data extraction was swift. Prosper had wrangled the data from internal systems which simplified this process. Since the dataset is sourced from a previously completed assignment, retrieving it simply required locating the dataset and documentation files in Windows. Which was done by hand since it’s not dispersed throughout the operating system.

1. Analysis Methods

1. The methods are tailored to fit a Descriptive Analysis (Part 1) and a Predictive Analysis (Part 2). This is from Part 1’s objective for finding correlation, and Part 2’s for predicting future outcomes in order to suggest actions for improving Prosper’s loan profitability. There is a broad scope of visualizations utilized. To identify correlation between numeric attributes, a Seaborn correlation matrix was initialized. This had a color scale to assist in separating the low and high R-values. Outside of this matrix, there were 3 chart types: univariate, bivariate, multivariate.

A univariate chart method provides a clear distribution or frequency for an attribute. Histograms visualized the distribution for continuous or discrete numerical values like Open Credit Lines and Credit Score. Plotted in them are 1st and 3rd quartile markings, for depiction of the Interquartile Range. Doughnut charts broke down categorical attributes comprised of values that represent an aggregated unit. Specifically, it was plotted for Term Length and Income Verifiable. A percentage was placed on each slice that clarifies a sum aggregation for that slice’s value frequency of the total. Bar charts show distribution of a categorical attribute, such as Prosper Score or Income Range. They have category values along the x axis and their counts on y axis. Each class member has a counter above its bar to provide an exact representation of distribution. If it’s ordinal, then they are correctly ranked.

Bivariate charts plotted distribution of 2 attributes, one categorical and one numerical, where the values of the categorical are compared amongst themselves. This method appeared as violin plots. I had class members as their own colors. The class attribute was kept along the x axis and the numeric was on the y axis. Violin plots had quartile markings similar to how a box plot does. Clustered columns are frequency distributions for two categorical attributes. There is an included colored legend for the 2nd class, and the 1st class members are on the x axis. Count intervals are on y axis. This setup enables easy distinguishing between classes, with rankings if ordinal. Heatmaps are also used as a bivariate method. A color scale helps identify the lower and higher cell densities. Similar to the others, categorical attribute is on x axis and numerical is on y axis.

A multivariate chart method is unparallel in its ability to have multiple dimensions within a single visualization. The prime type was a scatter plot. Two numerical attributes were plotted along the y & x axis. Then, colors and shape markers distinguished the 3rd categorical attribute that the numeric values represented. A legend was included for defining the color and shape for each class member. Blue circle was the first member, orange triangle was the second, and green rhombus was the third. An example is APR on y axis, Loan Amount on x axis, and Term Length as the represented category. A secondary type was used, Facet Grids. Similar to a scatter plot with the axis setup, but different with the representation of the categorical attribute. These consisted of many heatmaps – one for each class member. Every heatmap lies within the same grid, so there is a minimal amount of scrolling to view the data.

The final methods are tables and text. When comparing model accuracies and the prediction results against the true results, the model outputs are shown alongside each other in a Pandas data frame, essentially a table. These tables are additionally used for examining statistics of the dataset like standard deviation and quartiles. When there isn’t enough data, text is used. This method was enacted when scoring individual models or statistics & values of a single attribute.

1. Advantages & Limitations

Limitations are mainly contributed to the fact this was done with an open code platform and not a GUI application. This imposed a requirement of writing functions and code statements to do even the simplest of tasks. Visualizations sometimes required many lines of code for the customizing and setting up subplots within the plot figure. It’s substantially more complicated than when having clickable options or drag and drop features that exist in other environments. This technical method of analysis prolonged the duration. I had to spend time reading through manuals and documentation when constructing the models because there was no assistance, unlike those embedded in an application. Trial n’ error was paramount in configuring many of the code statements.

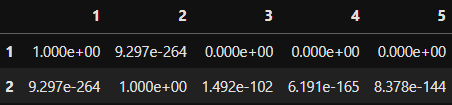
Advantages are present, even when considering these limitations. The benefit of a less structured, free-form environment is high customization. The charts, tables, and text outputs were created to my liking and specific to the intentions I have. GUI applications don’t offer the same creative freedom as the environment I analyzed in. Another perk was complete functionality for fine-tuning any model or code statement. Pre-built ones don’t always provide a comprehensive way to define parameters or options. It enabled granular level of detail over the commands I ran - at the cost of a slower pace. While prebuilt commands are simple to execute, they are limited in what they can output.

1. Step-by-step explanation
   * Step 1: This step regards verifying the input data for any chart, table, or text command. The dataset must be cleaned & assessed. The inputs should have a uniform structure and meet veracity guidelines. Determine if additional data is required to answer the research question or sub-problem being focused on.
   * Step 2: Collect the inputs within a single object, which should be a Pandas data frame. Field consists of a single value, columns headers signify what’s in the rows below them, and the records are unique entries. Ensure the inputs are complete and there aren’t missing entries or records. This satisfies the requirements to begin development of my output(s).
   * Step 3: Construct code building blocks to form a statement that takes the data frame and applies a function to it. This can be aggregation, returning a statistic, or some form of value retrieval. For text outputs, assign this function to a variable then use print(variable\_name) for displaying the desired output. For text, the process ends here.
   * Step 4: In situations of multiple outputs there are two options: charts or tables. For either, run the function as many times necessary to return a valid result. This may involve running it on multiple attributes from the input, creating derived variables, aggregation on class members for many classes, or summarizing the inputs with an assortment of statistics.
   * Step 5: These newly generated outputs should be stored in a dictionary format. The key represents its column’s header, and a value or list of values it contains are stored as that key’s value. Then use the Pandas DataFrame() function to transform the dictionary into a clearly labelled and organized table. For tables, display the table and end this process.
   * Step 6: Choose the attributes from the data frame that are needed to create an informative visualization. With the attributes selected, determine if a Multivariate, Bivariate, or Univariate is optimal given the dimensions of the selection. One attribute is an indication for Univariate, two are an indication for Bivariate, while three or more are an indication for Multivariate.
   * Step 7: Now write the code to set up that appropriate chart. Configure the parameters and included functions to customize the chart. Ensure a legend is included, both axes are labelled, proper title, fit y and x intervals/ticks, binning continuous attributes, pick a color scheme, and any visual aids like quartile markings or data point shapes based on the class member.
   * Step 8: Evaluate whether this finalized chart communicates the outputs effectively. If information is missing, consider adding subplots of the same chart type with various attributes or create a PairGrid(). If the information is congested, use a FacetGrid() to subvert data into other charts.
   * Step 9: In a markdown cell below the chart(s), identify the key findings and any relationships found. Describe the impact these have on my analysis. Answer the question: “what practical benefit does this insight procure?” If a relationship is exposed, investigate if it exists anywhere else in the dataset.
2. Results
   1. Model Outputs

A sample was made for equivalently populated groups, n size is determined by the group with lowest count. N size for test 1 & 2: 5. N Size for test 2 & 3: 6. Homoscedasticity and normality were evaluated through a combination of box plots and histograms. When the population distribution didn’t follow Empirical Rule or had a skew, along with abnormal shape, center or spread, normality wasn’t shown – indicating ANOVA couldn’t be used. This tactic was also performed on samples for substantiating similar distribution among groups, which is the main factor behind deciding on a Dunn or Conover Iman post hoc test. The models with parametric qualities were chosen for their improved precision, when possible.

For post hoc tests, Boolean values represent whether the test for a group pair is equal to or less than the alpha. This is done since legible P-values aren’t returned, they’re an unfinished equation.

They also return no test statistic. Example post hoc row without any Boolean modification:



Furthermore, the group sample for each test was plotted with both a box plot and histogram. These are compared to the population plots to validate there is no sampling bias. Group samples are representative of their population. For every test 1 – 4, there are 2 box and 2 histogram plots.

Test 1 – Kruskal Wallis: Sample of Credit Rating Groups for Loan Amounts

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of at least two groups.
* Statistic: 5526.157
* P-value: 0
* Alpha: .05
* Conclusion: With a P-value of 0, statistical significance is proven by Kruskal Wallis model, rejected null hypothesis. The test statistic of 5,526 is evidence of a moderately weighted significance, as one of lower strength would be a lesser value like 1,000. The box plots of each group depict medians of varying values, so this result aligns with the plots. Contradictory results would’ve appeared as a positive Kruskal Wallis but a similar median among all groups. This model was chosen for its robustness against non-homoscedastic and non-normality groups, hence the focus on median and not mean. Those traits paired with samples of equal sizes confirms this test is not only valid, but accurate.

Test 1 – Dunn Test: Sample of Credit Rating Groups for Loan Amounts

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of groups.
* Statistic: N/A
* Alpha: .05
* P-adjust: Holm-Bonferroni method.
* A screenshot of a computer

  Description automatically generatedP-values:
* Conclusion: Dunn Test was chosen due to the group distributions having characteristics of: right-skewed, unimodal, double peaks (besides sample 0 with one peak), and asymmetrical. Significant differences are detected between groups other than for group 4 (Very Good) and group 5 (Excellent). This is evident within the sample box plot from these two groups having the exact same mean, and the others ranging in Interquartile range and median by up to thousands. Poor (~2,250), Fair (4,000) and Good (5,000) simulate a hierarchy where the better credit rating score increases in both qualities until Very Good, where they become static. The sole variation between the 2 higher groups and Good is a median difference of ~1,000. The Loan Amount distribution is similar otherwise. This variation intensity decreases as you go from Poor upwards. This box plot is further evidence for the results from my Dunn Test.

Test 2 – Kruskal Wallis: Sample of Credit Rating Groups for APR

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of at least two groups.
* Statistic: 10468.13
* P-value: 0
* Alpha: .05
* Conclusion: Reject null hypothesis. With a P-value of 0, statistical significance is proven by Kruskal Wallis model. The test statistic of 10,468 is evidence of a strongly weighted significance, considering how its double the statistic of test 1. The box plots of each group depict medians more dispersed along the y axis. Contradictory results would’ve appeared as the box medians having less deviation than of the previous sample. This model was chosen for its robustness against non-homoscedastic and non-normality groups, hence the focus on median and not mean. Those traits paired with samples of equal sizes confirms this test is not only valid, but accurate.

Test 2 – Conover Iman: Sample of Credit Rating Groups for APR

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of groups.
* Statistic: N/A
* Alpha: .05
* P-adjust: Holm-Bonferroni method.
* A screenshot of a computer

  Description automatically generatedP-values:
* Conclusion: Statistical significance is detected between every group. Conover Iman Test was chosen due to the group distributions having center, spread and shape qualities unique to that individual group. Significant differences are detected between every group. This is evident within the sample box plot from these two groups having medians that vary from ~.11 (Exceptional) to ~.27 (Poor). A hierarchy is simulated where the better credit rating score decreases in median and Interquartile Range. The largest difference in median exists between Good and Very Good. Distribution for Poor and Excellent are inverted. If you took the Poor box and rotated it 180 degrees, you would get the Excellent box. This box plot is further evidence for the results from my Conover Iman Test – as no 2 groups are alike in center, shape, spread, and intensity of difference.

Test 3 - Kruskal Wallis: Sample of Income Range Groups for Loan Amounts

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of at least two groups.
* Statistic: 1.56
* P-value: 0.91
* Alpha: .05
* Conclusion: Fail to reject null hypothesis, there is no statistical significance between the medians of the groups. While groups are homoscedastic, the distribution is a far cry from Gaussian. Right-skewed, almost tri-modal, and no bell curve prove this. The box plots show Interquartile Ranges and medians of the same value across all groups - couldn’t be more homoscedastic if they tried. Sadly, there is no test that caters to this non-normality besides Kruskal Wallis. The low statistic is a precise portrayal in the distribution of the groups given their next-to-no variance. Since the p-value is .91, and the alpha is .05, no Post Hoc test is required.

Test 4 - ANOVA: Sample of Income Range Groups for APR

* Null hypothesis: There isn’t a statistically significant difference detected.
* Alternate hypothesis: There is a statistically significant difference between the medians of at least two groups.
* Statistic: 1.239
* P-value: 0.29
* Alpha: .05
* Conclusion: Fail to reject null hypothesis, there is no statistical significance between the medians of the groups. Groups are homoscedastic and the distribution is semi-closely matched Gaussian. It has a loose bell curve, slight negative skew, and large n size of 1,455 for the central limit theorem. These traits met the requirements for ANOVA. The box plots show Interquartile Ranges and medians of the same value across all groups - couldn’t be more homoscedastic if they tried. The low statistic is a precise portrayal in the distribution of the groups given their next-to-no variance. Since the p-value is .29, and the alpha is .05, no Post Hoc test is required.

## Loan Outcome Models

In total there are 6 models: Random Forest, Decision Tree, Logistic Regression, AdaBoost, XGBoost, and K-Nearest Neighbor. They are all evaluated using the same protocol. There are 3 datasets: training, evaluation, and testing. A cross evaluation score for accuracy is done then is compared to the accuracy from the training set. Additional metrics include precision, recall, and f1-score. Now from the evaluation set, this is measured with a cross evaluation prediction. These results are compared to real predictions output by each model. A final round of predicting is done. There is no cross evaluation involved with this one, only having the models predict on the test dataset then compared to the true answers I extracted from that test set. True answers were stored separately, so the models don’t have the information necessary to cheat. Cross evaluation was done with a 5-split stratified k-fold to preserve class distribution. The average from these 5 k-folds is taken to compare training accuracies.

The exact same processes were used to develop every model outside of the algorithm used and tuning of main parameters. Algorithms and scoring functions initialized from Python libraries of Scikit Learn and XGBoost. They’re all categorized as supervised learning classification. Metrics were rounded to the tenths decimal place for percentages. Parameter tuning was finished after feature selection (reduced noise) and when the cross-evaluation tests were aligning with the respective metric. Columns input: Income Range, Income Verifiable, Credit Score, Borrower Rate, Loan Original Amount, Open Credit Lines, Recommendations, Employed, Full-time, Other, Part-time, Retired, Self-employed.

A noticeable impact on these results was the inclusion of 3 predicted outcomes: defaulted, charged off, and completed. This was adjusted to 2 from defaulted and charged off’s similarity. When defaulted and charged off are merged into a single “defaulted” value, the models were able to surpass previous scores.

A screenshot of a graph

Description automatically generatedTraining Scores:

As depicted in the above table, the average from 5 cross evaluation scores was precise to the training accuracy score. This is evidence of minimal or no overfitting taking place. For each column Random Forest has the highest score, then XGBoost. Decision Tree has the lowest with KNN an overall point above. Research studies I inspected were able to get relatively higher training scores into the 70%’s & above, but in my attempts the models were overfitting when hitting scores that high. My reliance on continuous numeric attributes weren’t as optimal for training since they were preventing the models from generalizing. Logistic Regression had the best precision score with 68% - indicating that when focusing on the quality of the model’s true positive performance, it takes the crown. In terms of performance against all positive instances within the dataset (recall), Random Forest takes the crown. The F1-score was a tie between XGBoost and Random Forest, which is unsurprising given their similarity across metrics. XGBoost is inherently a modification of the Random Forest algorithm. These scores prove that the models will be capable of predicting future loan outcomes accurately, supporting my secondary research question.

A black screen with white text

Description automatically generatedEvaluation Predictions:

This table is contrived of prediction results based on the evaluation set for each model. There are 2 entries for each model: 1 for the cross-evaluation prediction and 1 for the real evaluation prediction. It exemplifies that my overfitting conundrum was resolved. The rather complex algorithms of Random Forest and XGBoost performed the worst when predicting by a margin of thousands off from the next lowest score – Logistic Regression. These models are failing to generalize from training and instead obtain higher metrics based on memorization. Since the input consists of a restricted number of features, this is affecting the abilities of complex models while boosting the effectiveness of less complex. The lesser models are Decision Tree, KNN, and AdaBoost. AdaBoost was the farthest off from the true answers in this lineup, coming in at barely different predictions is KNN and Decision Tree. These 2 models are off by tens instead of hundreds and even thousands like the others. It’s apparent that training doesn’t totally reflect the models’ performance to unseen data. When referencing the training table, the rankings would be swapped. I suspect KNN or Decision Tree is going to outperform other models from the closeness of evaluation predictions for both the cross-evaluation and true answers. The KNN and Decision Tree scores are supportive evidence of my secondary research question.

A screenshot of a computer

Description automatically generatedTest Predictions:

This table is composed of model predictions based on the test dataset. I ran multiple tests and KNN consistently outperformed Decision Tree by tens to around a hundred. The scores in this table were commonplace throughout my testing. Logistic Regression was unique in its tendency to output nearly identical Completed and Defaulted predictions. XGBoost and Random Forest were always convoluted with false positives. AdaBoost was overall a couple hundred predictions off from the true answers. Models boasting superb performance were Decision Tree and KNN. Due to a marginal difference of 50-75, or less every test run, KNN confirms my secondary research question as a – yes, new applicants can have their future loan outcomes predicted. This difference expressed as a percentage is .007%, meaning that KNN is correct 99% of the time for unseen data.

Lastly is the imputation model for Income Range and Employment Status. It runs on a Logistic Regression supervised learning classification algorithm. The training and testing accuracy were both 36%, which is impressive considering the model is bi-feature and multi-class output. If random sampling was to dictate these values instead, the correct value would be imputed 2.3% of the time. The success metric here was to get effectiveness that surpasses random sampling, and it has done that more than ten-fold.

* 1. Practical Significance

There was no statistical significance regarding the Income Range tests which does carry practical significance. Tests 3 & 4 are influential in decision-making because of this absence of statistical significance. It represents Prosper not taking appropriate measures to guarantee affluential demographics a more favorable loan offer. Wealthier individuals present higher returns since they are certified to have a larger monetary presence. This return comes in the form of more loans taken out of larger amounts, recommendations they can provide their affluential friends resulting in extra business, and less risk of defaulting from their financial stability. Other financial institutions who are taking advantage of these truths will gain customer favor from offering improved loan conditions. While the other banks are capitalizing on the financial status of these individuals, Prosper feigns ignorance and hands them clear reasons to choose a competitor bank – in essence it’s voluntarily losing profit.

A screenshot of a black and white screen

Description automatically generated

Across all groups, from $0 to $100,000+, applicants are just as likely to receive conditions. This table should depict a hierarchy where APR decreases and Loan Amounts increase when going up the ranks of Income Ranges. These statistics come from the population and aren’t restricted to the test samples. It’s evidence for the mishandling of customer’s applications on Prosper’s part.

For tests 1 & 2 statistical significance was proven, besides the difference in Loan Amount between Very Good and Exceptional; likewise practical significance is also apparent. These tests demonstrate what properly allocating loan conditions looks like. The lower credit scores have higher APRs and lower Loan Amounts.

A screenshot of a black screen

Description automatically generated

In regard to the Income Range effect on Prosper’s profit margins, this is the opposite but beneficial. It signifies improved treatment towards the affluential crowd which translates into financial gain in the future – indicated by the depicted hierarchy. Customers in each Credit Range are less likely to jump ship and bring business to a competitor bank due to receiving market rates and conditions. They have no reason to carry on with business elsewhere. When they are in a position of needing another loan in the future, they are guaranteed to consider Prosper. This truth is mainly derived from them having a positive outlook on their loan experience and will prematurely attribute that positivity to their future loan. It’s a win-win situation. Customers have found a bank that gives them competitive market rate conditions and Prosper gains their respect when they’re loan shopping. Lastly, having the foresight to permit higher loan amounts to deserving customers accumulates profit. This happens because the loan amount increase outpaces the APR deduction resulting in profit gains.

* 1. The Overall Success

The criteria for success I established in my proposal have been accomplished. The 1st was to identify significant difference among the groups Income Range and Credit Rating. For each group, Loan Amount and APR was measured. The tests gave a positive for credit rating affecting both measurements, with income range affecting neither. A hierarchy exists within credit rating. Proving this gave cause to evaluate the associated profit. The average loan amount and APR for each group was used in calculating this.

A graph showing a credit rating

Description automatically generated

Good is shown to reap the highest earnings in both statistics. For average earnings, Very Good and Excellent are nearly as lucrative but they have lower populations which causes Fair to overtake them in total profit. Poor only has approvals for minimal loan amounts, that along with this group being having only double the population of Excellent sends them to last place. With Income Range not having any significant difference detected, a chart isn’t appropriate. There is no added value Prosper is providing to affluential individuals, only reputable credit ratings.

The 2nd success metric was devising a model that was pronouncedly more accurate than random sampling for imputing Income Range and Employment Status features. The dataset had a vast amount of these values as NULLs, and I intended to rely on these features for my loan outcomes model and visualizations. Leaving them as NULL or imputing mode wasn’t as viable. Mode imputation wouldn’t have been very effective given the somewhat equal distribution among the 2 features. My Logistic Regression model impute achieves so with an accuracy of 36%, where random sampling would’ve been 2.3%. This is the enormous improvement I was looking for.

The 3rd success metric was producing loan outcome model scores within similar vicinity of the research studies I viewed. These would range from 70% - 90% for F1-score, recall, precision, and accuracy (V, V., A.C, R., K N, V., & G, A. | 2023) (Al Mamun, M., Farjana, A., & Mamun, M. | 2022) (Dansana, Prasad, Mishra, & Patro. | 2023). I employed the same metrics during in-processing to draw a comparison in post-processing. While my scores were on average 10-20% lower, the range of features in my dataset are primarily numeric and based on credit factors. Other research studies relied on customer demographics. This relationship led to some models overfitting, which limited the data each model had to work with. If I had access to customer demographics to base my dataset off of then the 10-20% gap could’ve been closed. Within the context of this problem, the scores are within close enough range to label this as a success.

The 4th success metric was providing a model as a solution to Prosper’s approval process. When deciding if a loan should be approved or not, there is a hefty amount of risk associated with the wrong answer. It can enable a customer to default and lose Prosper the entire loan amount potentially, or vice versa a customer who will pay the loan off is denied and that profit is not absent from Prosper’s earnings. After rounds of testing, one of my models can predict the outcome of new applications with an accuracy of 99.4%. That’s a massive change from Prosper’s accuracy of 69% and the random sampling accuracy of 50%. This outcome decides that my model is nearly perfect in terms of effectiveness and can be used to guide the approval process at Prosper.

1. Conclusions
   1. Summary

The models developed are capable of accurately predicting loan outcomes to reduce potential risk. Optimizing Prosper's loan approval process is proven as more than just possible, it’s nearly perfect given the model’s accuracy of 99.4%. This minimizes applicant risk based on an individual’s qualities, leading to increased profitability and customer satisfaction. Implementation of this model will greatly benefit Prosper in all financial avenues. The predictive analysis this entailed has provided a deployable and compatible solution for any newcomer customer. The second part of descriptive analysis has substantiated proof of credit rating influencing various loan factors, mainly loan amount and APR. It’s a positive that such a relationship does exist, and it rewards Prosper and customers. Retention, satisfaction, and repeat business are more probable when these conditions are allotted based on the individual customer and not at random. The timelines and milestones were routinely conquered while remaining within budget. Resources have been adeptly applied in producing a modern solution for the research questions. The 4 metrics of success all indicate this project lives up to the promised results. Outside of predicting loan outcomes of future customers, necessary and often overlooked data is able to be imputed for precision within any analysis using Prosper’s data. It is guaranteed to be quite handy for a variety of scenarios. I explored the relationship of Prosper Score and other attributes such as Credit Score and Open Credit Lines, but no relationship was found other than Debt To Income Ratio having a negative correlation. This work stands as a testament to the power of data analysis and machine learning in improving financial services, underscoring the importance of innovation in the sector.

* 1. Storytelling

The visualizations display true information and are staples in how effective storytelling is within my presentation. The profit charts based on the loan outcome model give a level of depth in showing stakeholders the importance of my solution. These are figures known personally to them and are memorable in the data they contain. While specific model accuracies or even the metrics of my best model aren’t significant to them, having the impact relayed in a familiar manner instigates a deeper understanding. It’s a line of communication that is otherwise unachievable via charts littered with technical jargon or irrelevant figures. They are interested in findings being applied to the fundamental principles known personally. Statistics of standard deviation or distributions aren’t as influential compared to less complex statistics like average or totals, which I take advantage of. Leaving an impression that I’m a person who understands their point of view and builds rapport through familiarity in what my charts convey is beneficial. The non-granular details these visualizations focus on are the gateway between my findings and their application of those findings. The charts were picked for simplicity and not complexity. Pie, bar, scatter plots, and tables are flawless vessels for me to bridge the gap between raw data and conveying valuable information. Highlighting the key takeaways for each chart are placed in a visually pleasing format via PowerPoint presentation. It has the functionality to insert the types of charts I need, while also supporting the dialogue with slide room for short sentences or bullet points. This file format is easily accessible and can be referenced by audience members on their own time.

I have intent to demonstrate the uncertainty a customer faces by assuming their perspective and the fallout of dealing with unfair treatment from improper loan conditions or an approval denial. The monetary strain of having too many of these relationships and the dissuasion that occurs from customers avoiding potential business with Prosper is the strategy. Seeing and feeling these incorrect loan approvals as human emotion is vital. The connection I’ll obtain by depicting the harsh reality of both applicants and Prosper will be suavely done throughout the PowerPoint presentation.

* 1. Recommendations
     + For incoming loan applicants, submit the required customer data into the machine learning model for determining an approval or denial.
     + Customers with a lower income in the range of $0 - $24,999 or in an upper range of $75,000+ should be provided loan conditions that factor this in.

**Lower range:** max of $5,000 amount and above .25 APR.

**Upper range:** max is dependent on their income and under .20 APR.

1. References

Ml/ai, I. V. T. I. D. (2023, August 20). *Model Complexity in Machine Learning - Ishan | Virginia Tech & IIT Delhi | ✍️ ML/AI - Medium*. Medium. <https://ishanjainoffical.medium.com/model-complexity-explained-intuitively-e179e38866b6>

Wijaya, C. Y. (2022, January 6). *CRISP-DM Methodology For Your First Data Science Project*. Medium. <https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c>

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>.

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>

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>

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>