Nicolas Bolanos Parra

Lead Scoring Project

PREDICT 490 Course Project

Introduction: Brackground

- About the Company:
  - Founded in 2008, the Company is a private money lender for real estate investors who need quick access to capital to finance residential investment transactions.
  - **Very specific niche industry.** Financing is provided for residential investment property purposes only.
    - The Borrower may not occupy the property.
    - The Borrower must have one of the following investment strategies:
      - Fix and Flip: Purchase a distressed property, repair all the damages and sell it at retail price.
      - Fix and Rent: Purchase a distressed property, repair all the damages and hold it as a rental property.
      - Refinance and Sell/Rent: Repair a property the Borrower owns in order to sell it or rent it out.
  - **Borrower qualification.** Loan approvals are based on whether the investment makes financial sense and whether the Borrower is skilled enough to complete the project. Credit score, DTI and other metrics used in conventional financing are not determinants for qualification.

Introduction: The Issue

- Approximately 70% of new leads come from the Internet.
- Approximately 10% of new leads are qualified prospects.
- Approximately 60% of qualified prospects turn into closings.
- Sales representatives spend 25% of their time speaking to non-qualifying or low value leads.
- Increasing marketing costs in an attempt to keep up with sales volume expectations.

Proposed Solution: Lead Scoring

- In an effort to optimize sales and marketing resources, a lead scoring system is proposed to allow sales representatives and marketing managers prioritize opportunities.
    - Sales representatives can prioritize their time and effort by speaking with the most valuable prospects
    - Marketing managers can identify and target high opportunity markets.
    - Provides opportunity to nurture prospects for future business.

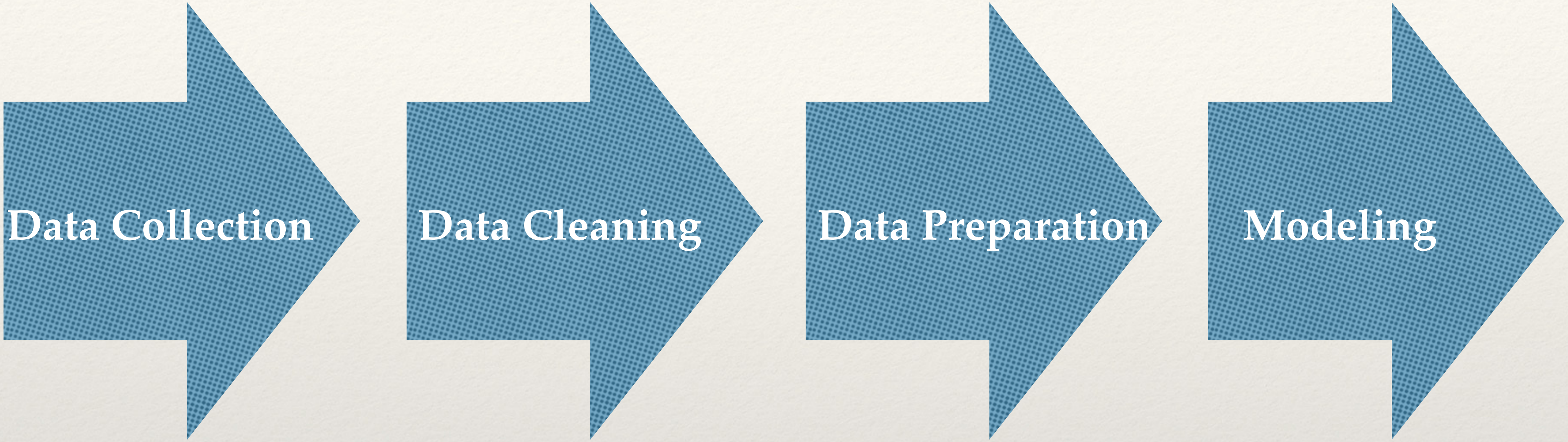
Project Scope

The proposed solution...

- Is not intended to replace the current underwriting criteria for loan approvals.
- Is not applicable to commercial or new construction investment projects. These are more complex and their success depend on additional factors.
- Does not consider some factors affecting the value of a lead such as customer lifetime value or risk of loan default.
- Is designed for loan amounts of up to $500,000. Larger investment projects, while possible, require third parties and can introduces additional variables not considered here.

Process and Methodology

Steps to Develop the Model



Data Collection

Data Cleaning

Data Preparation

Modeling

Data was extracted from multiple sources and compiled into a single file using Python. (all_transactions.csv)

This is where data format issues and missing values were handled. These data issues were handled programmatically using Python

Data was prepared and made suitable for modeling using SAS. This is where outliers and some missing values were handled.

A probability-severity model was used to score the leads. SAS was used in this step.

Data Collection

- Data was collected from multiple internal and external sources.
- **Internal Data.** Collected from two main sources:
  - Forms filled out by customers during previous successful and unsuccessful sales.
  - Entries on the CRM by sales people.
  - Loan pipeline
- **External Data.** Collected from third party sources.
  - Zillow
  - Trulia

Data Sources: Internal

Borrower Information

Variables	Description	Theoretical Effect
Years of Experience	Number of years of experience in the Real Estate Investing Industry.	The more experience the higher the likelihood of closing a loan
Number of Completed Properties	Number of completed properties at the time of application	The more properties completed the higher the likelihood of closing a loan
Repeat Borrower	Whether the prospect has done business with the company in the past	Current clients know the process of closing a loan and will know what to expect, thus increasing the likelihood of closing a loan.
Cash Reserves for Investing	Verifiable cash reserves to finance initial carrying costs of the loan.	The more cash available, the higher the likelihood of loan closing.

Data Sources: Internal (Continued)

Investment Property Information

Variables	Description	Theoretical Effect
Property Purchase Price	The purchase price of the property in distress	Low purchase prices relative to the after-repaired value have high profit margins and are likely to close. However, in general, low purchase price are less valuable to the lender.
After Repaired Value (ARV)	The value of the property after all the needed repairs have been completed	Loans for properties with high ARV are more difficult to close (less likelihood of closing), but are much more valuable to the lender.
Repair Credit Line	The estimated dollar amount provided to the borrower to complete property repairs.	The more repairs required on a property relative to the purchase price, the less likely it will close. The case may be more valuable to the lender if the prospect is experienced enough for the project.
Bedrooms	Number of bedrooms	Properties with 2 or 3 bedrooms are easier to flip or rent and will likely close.
Bathrooms	Number of bathrooms	Properties with 2 bathrooms are easier to flip or rent and will likely close

Data Sources: Internal (Continued)

Transaction Information

Variables	Description	Theoretical Effect
Square Footage	Heated square footage of subject property.	The larger the property the higher the loan amount is likely to be. It may have a negative effect on the likelihood of closing.
Closing Date	The date of loan closing	During some times of the year loans are more likely to close and are likely to be more valuable to the lender
Loan Amount	The dollar amount of the loan	Target Variable
Closed	Whether or not the transaction was successful	Target Variable

Data Sources: External

Market Health Information

Variables	Description	Theoretical Effect
Market Health Index	Zillow's real estate market health index	Economically stable or growing markets have a positive effect on the outcome of the investment and are thus more likely to close.
Median Sales Price & Median Sales Price/ SqFt	Zillow's month-to-month median sales price (and sales price/SqFt) estimation (by zip code)	Properties are appraised based on comparable properties within the same area. Higher median sales prices indicate high valued transactions.
YoY	Year Over Year appreciation or devaluation of properties in the area for the current year.	Loan approvals are contingent to appreciation prospects of the neighborhood.
Days on Market	Average number of days for a property to sell in the market (by zip code)	Loan approvals are contingent on the marketability of the property. Properties in areas that sell quicker are likely to be approved more often.
Median Sales Price per Square Foot	Zillow's month-to-month median sales price per square foot estimation (by zip code)	Properties are appraised based on comparable properties within the same area. Higher median sales prices indicate high valued transactions.

Data Cleaning, Collection and Preparation Issues

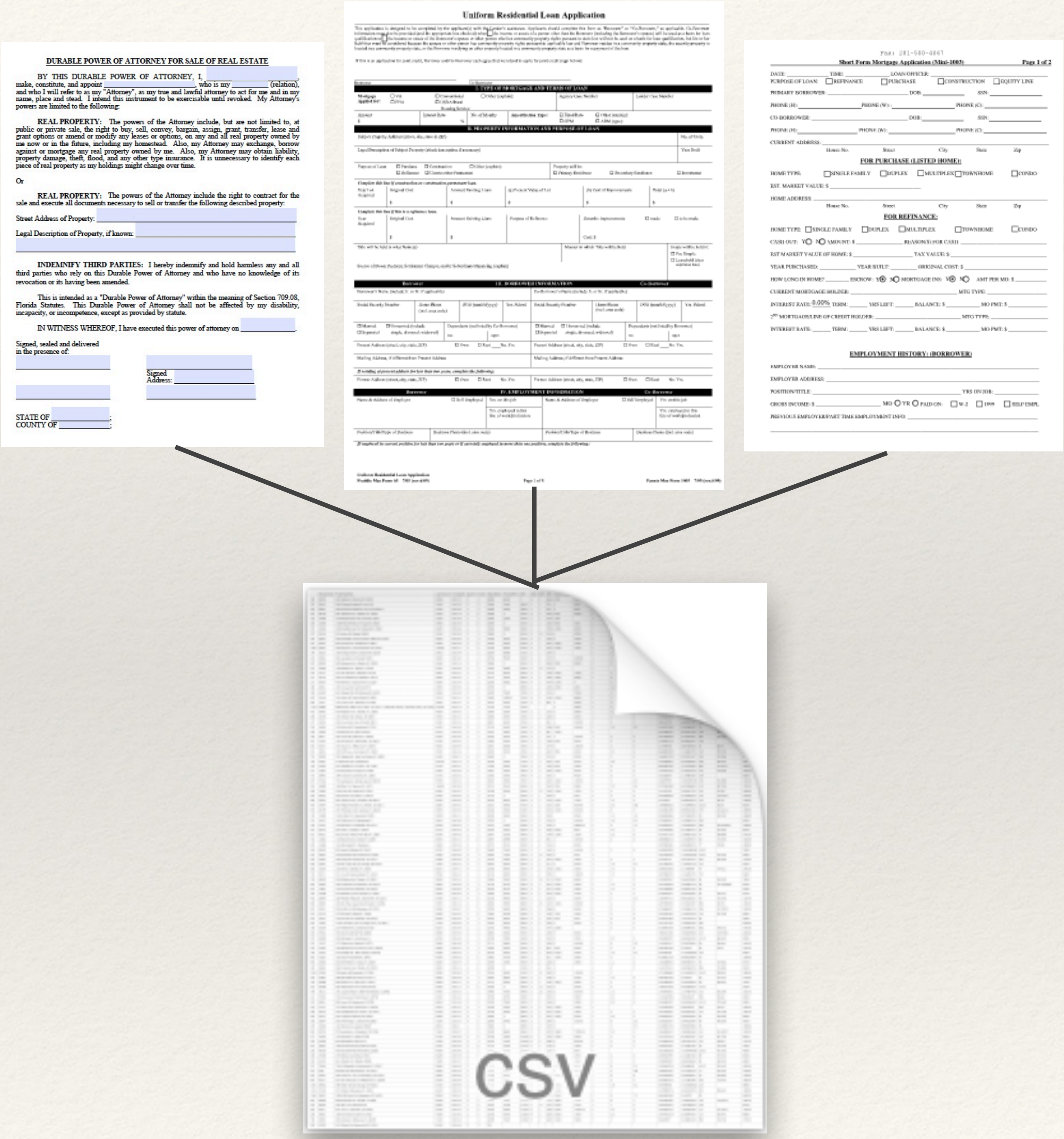
- There were three main data issues throughout the modeling process:
- Data Scarcity and Missing Values
- Data Format
- Outliers

Data Scarcity and Missing Values

- Data is collected internally via PDF forms. Borrowers may fill out these forms on their computer but some opt for filling them by hand, preventing the extraction and parsing of some data.
- Some PDF forms must be locked and signed by the Borrower, preventing the extraction and parsing of some data.
- Fillable data collection documents were not available during 2011 and early 2012.

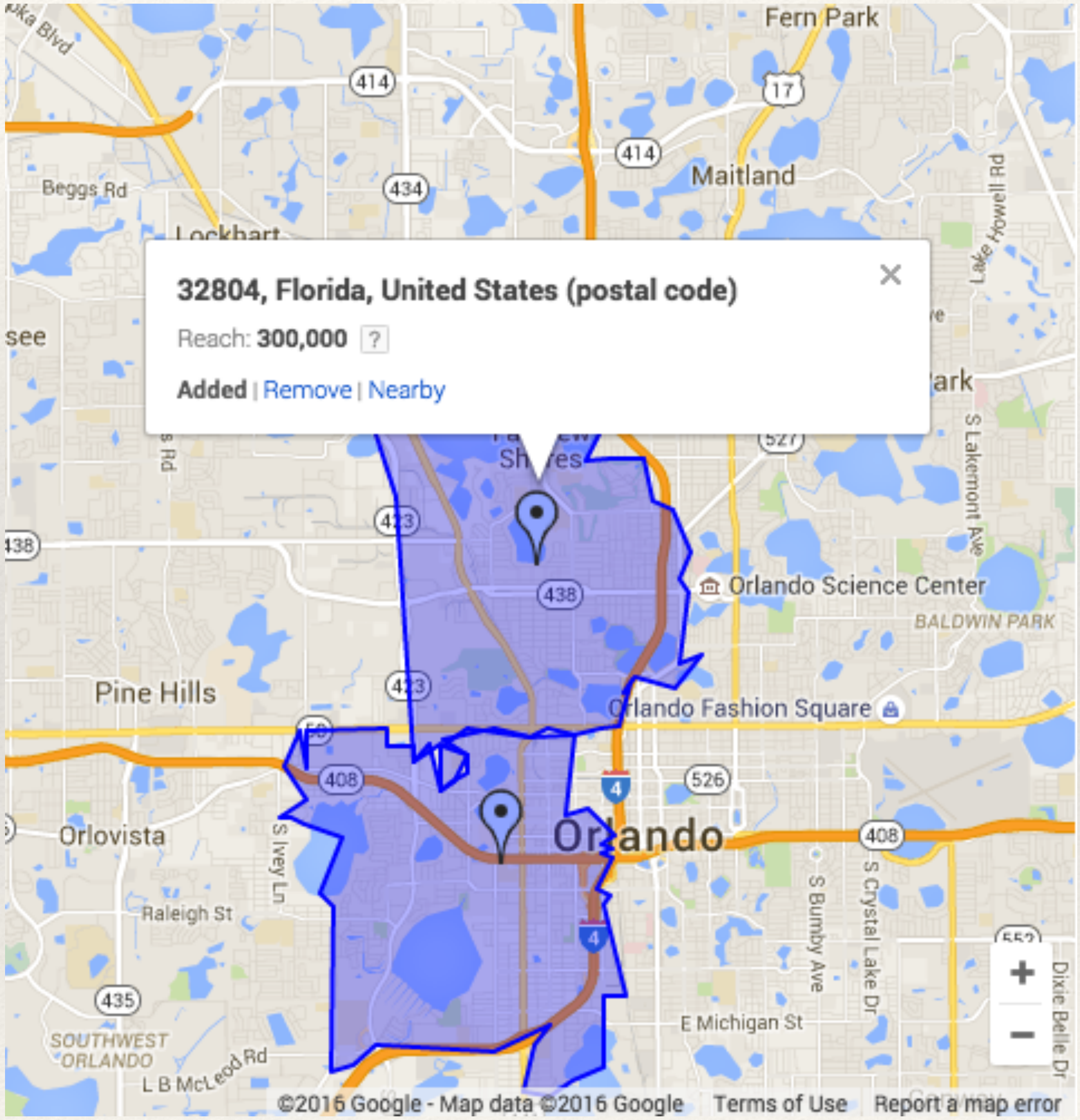
Handling Data Scarcity and Missing Values

- Data scarcity and missing values were handled with Python during the data collection process and with SAS during the data preparation process.
- There is some degree of data redundancy among the data collection documents used internally. **Multiple documents were collected and compiled for each observation**, reducing the number of missing values.
- Missing values related to the subject property were complemented with external data sources. **A web data extraction system was built to search and scan missing values that could also be found in public records**, such as the number of bedrooms, bathrooms and square footage of a property


DURABLE POWER OF ATTORNEY FOR SALE OF REAL ESTATE
Uniform Residential Loan Application
Short Form Mortgage Application
CSV

Handling Data Scarcity and Missing Values

- Most of the market information was found using a lookup function in Python. However, for some zip codes the data was not available. Using the pyzipcode package for Python, nearby zip codes were found, which were used to estimate market information.



32804, Florida, United States (postal code)

Reach: **300,000** ?

Added | Remove | Nearby

Handling Data Scarcity and Missing Values

- During the data preparation process in SAS, the remaining missing values were imputed using linear regression (regressing the variable with missing values on other available variables) or simply using the median value.

Data Format Issues

- There were slight differences among redundant data within the forms. Ex: Some forms asked for years of experience in months, others in years. This made the data compilation process difficult.
- The variables related to the borrower (years of experience, number of completed properties and whether they had completed a loan in the past) were only found in unstructured data formats.

Handling Data Format Issues

- Differences among redundant data within forms were handled programmatically in Python using the pandas and numpy packages.
- Unstructured data were handled programmatically using regular expressions in Python (for example by looking behind words like "years" to extract the borrower's number of years of experience).

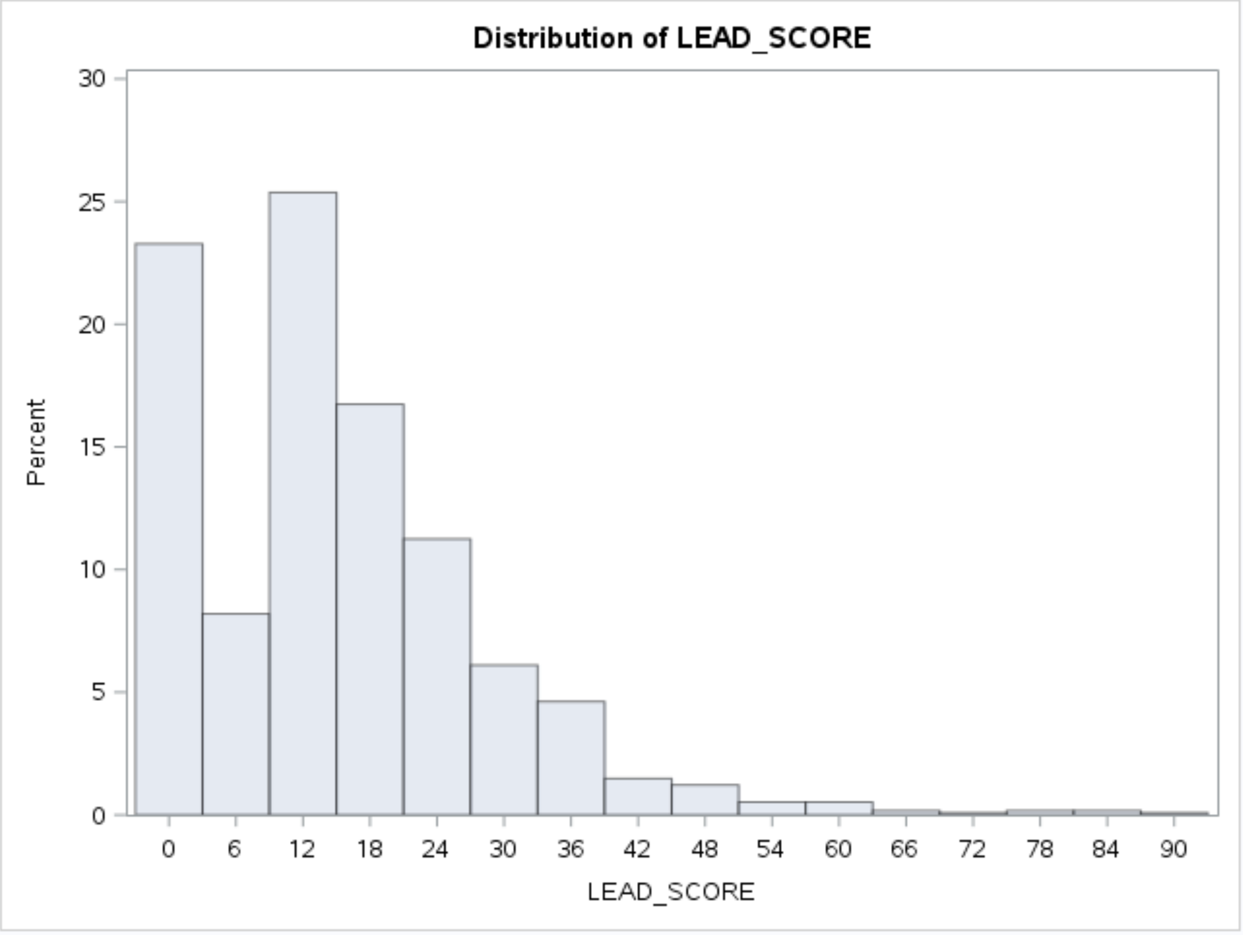
Handling Outliers

- Outliers were handled in SAS during the data preparation stage.
- Some observations containing outliers were deleted because they were outside of the project's defined scope.
- Variable transformations were used to mitigate the effects of outliers in the regression models.

The Model

- The value of a lead is contingent on the likelihood of closing a loan.
- Provided that the loan closes, the value of a lead depends on the dollar amount of the loan.
- A probability-severity model was used to model this situation, where both the probability of closing and the total dollar amount are considered to calculate the expected value of a lead:

LEAD_VALUE = PROBABILITY OF CLOSING * LOAN AMOUNT



Observed distribution of the Lead Score (to a 0-to-100 scale)

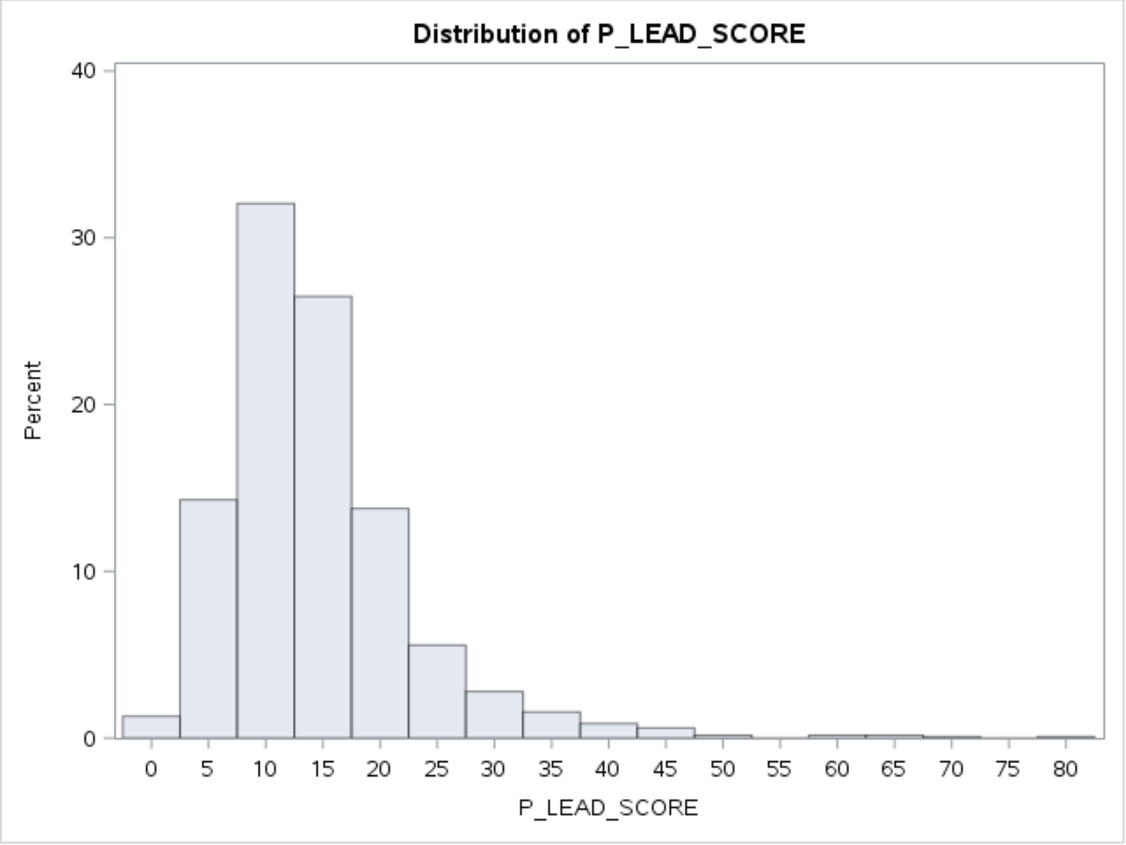
The Model: Scoring Function

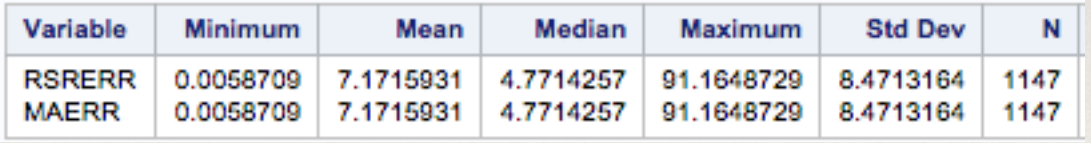
```
/*Scoring Function (Assuming no data is missing)*/
data SCORE;
set TEMPFILE;
P_LOG_LoanAmount = 1.42308 +
                        0.33721 * IMP_LOG_ARV +
                        0.00000055 * IMP_PurchasePrice +
                        0.000000739 * MedianSalesPrice +
                        0.00003807 * IMP_SqFt +
                        0.00049456 * IMP_MedianSalesPriceSqFt +
                        0.000000000867 * ClosingDate +
                        0.02303 * IMP_Bath +
                        0.01239 * IMP_Beds +
                        -0.01675 * IMP_LOG_EstimatedRepairs +
                        -0.02629 * M_CompletedProperties +
                        0.1051 * M_Bath +
                        -0.03015 * Quarter1;

P_LoanAmount = 10 ** (P_LOG_LoanAmount);
```

```
if(P_LoanAmount > 540000.00 or P_LoanAmount < 15960.00) then do;
      P_LEAD_SCORE = -1;
end;
else do;
      LOGIT_TEMP = 1.8665 +
                  0.9612 * RepeatBorrower +
                  0.0838 * MarketHealthIndex +
                  -0.1421 * IMP_YearsOfExperience +
                  0.00896 * IMP_CompletedProperties;

      ODDS_TEMP = exp(LOGIT_TEMP);
      PROBABILITY_CLOSING = ODDS_TEMP/(1 + ODDS_TEMP);

      LEAD_VALUE = PROBABILITY_CLOSING * P_LoanAmount;
      /*Scale the lead score to a 0 to 100 range*/
      P_LEAD_SCORE = LEAD_VALUE/(540000.00-15960.00)*100;
end;

run;
```

Outcome

- Using approximately 900 observations the model was created and evaluated on a sample of 1147 observations.
- The distribution of the predicted lead scores (P_LEAD_SCORE) has a mean of 10 to 15, which approximates the observed LEAD_SCORE distribution
- The model does not seem to account for the zero-inflated-portion of the observed variable.
- Some variables did not conform to the theoretical considerations: The number of properties completed by the borrower seems to affect negatively the loan amount and the borrower's number of years of experience seems to affect negatively the likelihood of obtaining a loan. It is possible that some variables are missing, or that there are data quality problems.
- The Mean Root Squared Error is approximately 7.17 which is the average deviation of the predicted value from the observed value.





Variable	Minimum	Mean	Median	Maximum	Std Dev	N
RSRERR	0.0058709	7.1715931	4.7714257	91.1648729	8.4713164	1147
MAERR	0.0058709	7.1715931	4.7714257	91.1648729	8.4713164	1147

RSRERR = Root Squared Error
MAERR = Absolute Squared Error

Outcome (Continued)

- **Model Strengths:** The model considers the latest and most localized real estate trends to evaluate the value of a new lead. It requires little information from the prospect, which is desirable within the context of lead generation (many users are not willing to provide much information on their first contact with a company)
- **Model weaknesses:** There are evident issues with prediction accuracy (more in relation to probability of closing predictions) that stem from the fact that the data was not readily available, organized into proper format, or it was missing altogether. In each iteration, the strategies presented here were implemented to mitigate this problem were used, and the results consistently improved. However, it is advisable to continue to refine the model once sufficient and reliable data has been collected.

Business Impact

- Although the model has not yet been tested there is great opportunity to do so in a production environment:
  - The model can be integrated into the online advertising strategy of the company, thus targeting areas where higher lead scores are observed.
  - The model can be seamlessly integrated into the sales process. Using the data collection and cleaning code and the lead scoring function, new leads can be scored on-the-fly and submitted to the CRM together with the new lead data. There would be no additional effort required by the sales person.

Next Steps

- More, complete observations are required to continue to refine the model.
- Other possible variables not considered here but that are relevant to the model are:
- Prospect Occupation: Realtor, Real Estate broker, General Contractor, Real Estate Entrepreneur, etc.
- Prospect Credit Score: It does not determine loan approval but can also be considered.
- Property Features: Whether it has a 2-car garage, vaulted ceilings, whether it has a pool, etc.
- Property Comparables Information: Information such as appraised value of comparable properties within 5 miles of subject property.

Next Steps (Continued)

- It is clear that further refinements are needed. However, there is currently no mechanism in place to gauge the value of the leads the company generates, other than the sales people talking directly with the prospect. A small change towards a more analytical sales function can make a substantial difference in the long run in time effort and the bottom line.

Conclusion

- First experience applying analytics to sales and marketing.
- Handling data issues was the most difficult part of the project, but it was also an incentive to find creative solutions to common problems.
- There was a substantial amount of programming with Python in this project, which provided excellent opportunities for learning the language and wrangle difficult data.
- Data issues also provided insights about how data should be managed within the company. To facilitate future predictive analytics work, the data should be stored in a central location where the data can be retrieved without the need of cross-referencing observations.