# Team B: Aleators

Professor Day Yi

APAN 5900

28 June 2020

# Research proposal: Introduction & Methods

# I. Introduction

## ORGANIZATION OVERVIEW

Spruce is a title and escrow company whose mission is “to make real estate amazing for everyone involved”. For escrow services, Spruce acts as a neutral third party that helps coordinate transactions between homeowners and their lender or real estate institution. For the title side businesses, Spruce underwrites title insurance which provides coverage against lawsuits involving the insured property.

## INDUSTRY RESEARCH

In 2018, the post-merger title insurance industry shrank from a "Big Four" to a “Big Three” oligopoly, changing from a moderately to a highly concentrated market level. Under the oligopoly, consumers find it difficult to decide over which title insurance to choose due to unfamiliarity with the products. Therefore, the government should take action to educate consumers and strengthen the regulation of the market.

Title agents and real estate professionals are aware that fintech is transforming the industry for faster and more efficient transactions. Therefore, the professionals are anticipating three key innovations: Secure collaboration and communication portals contributed 45%, E-Closing and remote online notarization contributed 34% and chatbots contributed 18% (MarketWatch). Once the innovation is fully implemented in the market, the industry becomes less concentrated and more competitive on price, thus the insurance premiums and service charges will fall.

## BUSINESS CHALLENGE

To analyze the market competition, SWOT analysis can help Spruce understand their competitive positioning in its business environment and analyze what contributes to the success of Spruce. Spruce’s SWOT Analysis is as follow:

|  |  |
| --- | --- |
| Strengths:   * Better digitalization improves the accuracy and efficiency of transactions; * Two streams of revenue provide a clear value proposition and growing transaction volume; | Weaknesses:   * Property and information security problems cause confidentiality getting compromised; * Legal issues will occur when dealing with a different state’s cases; |
| Opportunities:   * Spruce has great technology advantages to reduce the probability of title frauds; * Customer interests are rising in New York state’s real estate market in recent years; | Threats:   * Price fluctuation will have an uncertain impact on profits of Spruce; * Macroeconomics will influence the demands of Spruce products and services. |

Currently, Spruce performs underwriting on a per-file basis by conducting a title search, the business problem is how to improve the accuracy of title insurance assessment. This project aims at using analytics and building a model for Spruce to predict title defects which will enable Spruce to make more informed underwriting decisions and improve its operational effectiveness.

## RESEARCH QUESTIONS AND HYPOTHESE

Based on the historical dataset we have, the first research question we raise is about the dataset quality since it is the foundation of the whole project. **RQ 1:** **Is the quality of this dataset good enough to predict title defects?** The null hypothesis we set here is the dataset is not in good quality. One possible way to test this research question is to compute the proportion of observations with title defects in the entire dataset.

If we reject the null hypothesis for the first research question, then we move forward to the second research question. **RQ 2: Which variables are most important to predict title defects?** The null hypothesis we set here is there is no significant correlation between Total Market Value and Title Defects. We plan to test this by correlation analysis of these two variables.

# II. Data Processing Methodology

## DATA DECRIPTION & CLASSIFICATION

Spruce collects the client’s information related to different dimensions of the property to analyze the relationship between those variables and title defects and build predictive models to facilitate future underwriting decisions. The dataset contains over 500k observations and 52 variables. The data are distinguished by three types: there are 47 Integers, 4 Floats, and 1 String.

Among variables, the Target Variable is dr\_Title\_defect\_ind, which indicates whether a title defect exists (1 is title defect, 0 is no defect). And The dataset shows 1,786 target records. We then divided the independent variables into three main categories: **1)** **The attribute of the property itself**, including entity structure like counts of the bedrooms and stories, property value, both assessed value and market value. **2)** **The attributes of the region/household where the property is located**. Data provided detailed demographics information such as employment and income level of the residents. Besides, the real estate market situation is reflected by average rent and house prices. **3)** Beyond the information about the property, it provides **existing & predictive defects**. Existing ones include the percentage of households with bankruptcy, mortgage past due, or severe derogatory status, while the risk scores predict the likelihood of these serious credit risks in the next 2 years.

## PRELIMINARY SCREENING

After walking through the dataset, we decided to reduce the number of variables to build more efficient models and reduce extreme complexity. First, based on our understanding of Spruce’s business and the project, we roughly dropped some variables to reduce the dimension. We keep the variables that we think might have an effect on the target variable and drop those that we don’t see direct relation with title defects. Next, after filtering out some features that seem irrelevant to the target variable, we divided the remaining 20 variables in the data frame into two groups to better see the correlations. Through the heatmap visualizations, we were able to see some features which have high correlations with the target variable, including dr\_Ownership\_Period, X\_PERC\_CDPD60, X\_Risk\_Score, and so on. We also see some variables which are highly correlated with other variables, such as X\_Bankruptcy\_Index, X\_Risk\_Score, and X\_Vantage\_Score (See Appendix A). Then we picked several features to explore more on their distributions for title defects. For instance, by doing visualizations for the ownership period, we found that the title defect tends to have a shorter ownership period than normal titles (See Appendix B).

## FEATURE SELECTION

After our initial exploration of the dataset using charts and plots, we plan to use feature selection methods to select the variables that we would be using to build our predictive model. There are 3 main types of feature selection methods: Filter Methods, Wrapper Methods and Embedded Methods. We have decided not to use filter methods for this project since it does not involve the process of model training. Stepwise variable selection methods and shrinkage methods such as Lasso and Ridge will be applied to find the best variables. All these feature selection models lie in the Scikit-Learn package in Python.

## DATA CLEANING

The next stage after feature selection is data cleaning. The data cleaning stage can be divided into 4 steps: 1) Renaming Variables. We will initially rename variables that currently have obscure names or abbreviations, such as “dr\_PERC\_VACUNIT”. We plan to change it to “Vacant\_Units\_Percentage” for easier understanding. 2) Removing outliers. Then outliers should be removed accordingly. We will compute descriptive statistical values such as median and quantiles to get an overview of data distribution. 3) Transforming data. Next, we need to convert text values to string numbers to be used in the model. For instance, “dr\_Value\_Change” contains 3 levels of data: INC, DEC and NC. We would convert them to 1, -1, and 0 accordingly. 4) Dealing with missing values. Variables with missing values can be addressed based on two categories, importance and missing rate. For variables with high importance and low missing rate, such as “TaxMarketValueTotal”, we plan to calculate missing values according to industry knowledge; for variables with both high in two categories, we plan to get extra information from sources that we trust to fill in missing values; for variables with both low in two categories, we can either compute the mean or just keep them as they are; and for variables with low importance but high missing rate, such as “dr\_Ownership\_Period”, we tend to drop them.

## MODELING

Since the target prediction is a binary outcome, classifier models turn out to be the best choice. We plan to use four classifier models: Logistic Regression, Decision Tree/Random Forest, KNN and SVM. Each of these methods provides us with robust prediction yet brings various challenges as well.

### *Logistic Regression*

Logistic regression is highly efficient and easy to interpret since the final outcome is expressed as probability and can be used for ranking instead of classification. However, one major limitation of logistic regression is that it assumes a linear relationship between the logit of the outcome and each predictor variable. For our case further interpretation will be a must.

### *Decision Tree/Random Forest*

Decision Tree and Random Forest require less data cleaning effort and offer clear visualization for results. However, trees are prone to overfitting problems and based on heuristics such as the greedy algorithm, so there is no guarantee for a globally optimal decision tree.

### *K-nearest Neighbors (KNN)*

KNN is a simple and audience-friendly algorithm, easy to understand with no assumptions, yet KNN struggles to produce the model as the numbers of variables grow. Also, the choice of the K value needs constant modification during the process. (Bronshtein, Adi)

### *Support Vector Machine (SVM)*

SVM can solve the dimensionality problem for KNN and it is memory efficient because it uses a subset of training points in the decision function (Ray, Sunil). But SVM takes up too much time to run the result so it brings computational inefficiency. Besides, SVM does not directly provide probability estimates and interpretation may be a challenge.

# III. Prospects

# Above are our ideal and preliminary plan for the project. However, chances are that some classifier models will not yield good results as expected. Under that circumstance, we will make prudent decisions based on what we see, including the possibility of bringing in other methods. All in all, our ultimate goal is to build the best solution and improve efficiency for Spruce.

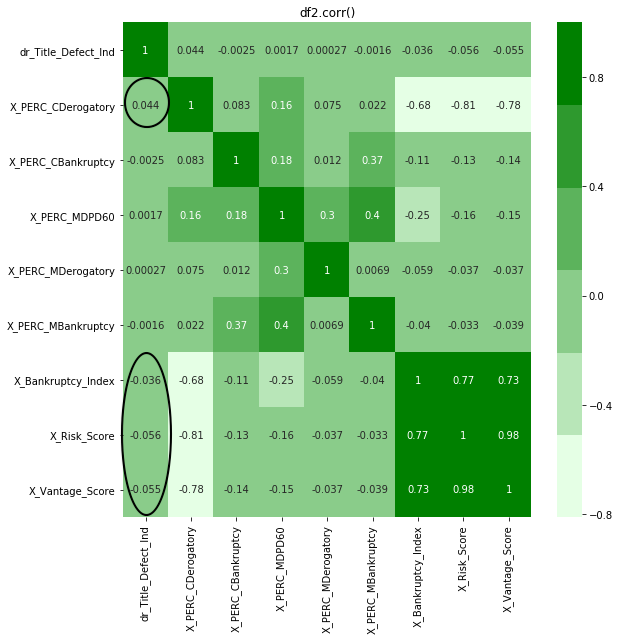
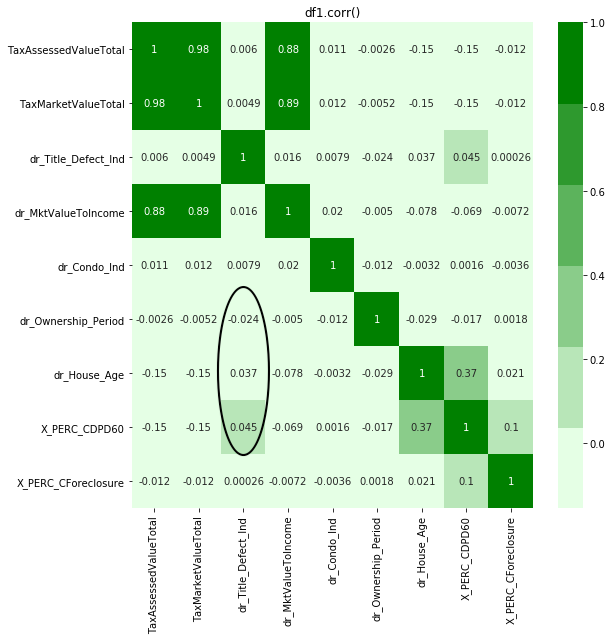
#### Works Cited

Bronshtein, Adi. *“A Quick Introduction to K-Nearest Neighbors Algorithm.”* Noteworthy - The Journal Blog. N.p., 11 Apr. 2017. Web. 26 June 2020.

MarketWatch. *Fintech Adoption Accelerating Among Real Estate Professionals*, According to First American Real Estate Sentiment Index. 27 November 2018.

Ray, Sunil. *“SVM | Support Vector Machine Algorithm in Machine Learning.”* Analyticsvidhya.com. N.p., 13 Sept. 2017. Web. 26 June 2020.

#### Appendix A



#### Appendix B

