Florida's Residential Single Family Home Re-sales Review

Predicting outcome of the single-family residential home same year re-sales

By Ruslan Dubas

Introduction

From 2009 to 2020, in Florida, there were on average 135,000 residential homes sales per year. This wealth of the information comes from the Florida Department of Revenue. The research into the data will be focusing on the single residential home sales (SFR) which were bought and sold again within same year. Will the investment create a loss or make a profit? The ability to accurately estimate positive or negative outcome, prior to investment, is an important task for all real estate market participants and stakeholders.

Analysis will be focusing on building explainable ("White-Box") binary classification models. The two classes which study will attempt to classify: Loss or Profit. Features importance will be measured by applying Logistic Regression, Decision Tree, and Random Forest algorithms.

Data Source: **Florida Department of Revenue (FDOR) Assessment Rolls from 2009 to 2020**

The following is the excerpt from the FDOR's website: "The Department of Revenue publishes assessment rolls in compliance with chapter 119, Florida Statutes. The files publicly available through the Department and county property appraisers do not contain confidential records, such as social security numbers and the records of property owners exempt from public records disclosure under section 119.071, Florida Statutes."

Brief Descriptions of the Data Source

FDOR publishes the NAL (*"Name, Address and Legal"*) files as comma-delimited files (with the file extension .csv) with field names in the header row. Each file contains 161 columns, detailed descriptions can be reviewed in the Appendix 1 – 2022 User's Guide. Each row represents a single real estate property parcel. 68 files for each year, from 2009 to 2020, were combined into one SQL Server Database Table for the analysis.

Gathering the Data

FDOR upon request provided the link to the NAL files. The main difficulty was to combine all 680 files (68 counties for each year from 2009 to 2020) into one and import it into SQL Server as one Database Table. The importance of such vast data gathering cannot be understated – once all the data has been combined the analysis of the complete population can be performed. Complete population, in this case, pertains to the Single Family Residential (SFR) home same year re-sales.

All SFR homes labeled as ‘001’ in the Use Code field. The nature of the transaction captured in the *Sales Qualification Code* Field. Multi-parcel (‘05’), corrective deed (‘11’) and transfer between relatives (‘30’) sales were removed from the data set. Detailed descriptions of all fields could be found in the Appendix 1 - 2022 User’s Guide.

Graphing All Florida SFR Home Sales from 2009 to 2020

The graph below shows counts of all Florida SFR home sales from 2009 to 2020. There were on average - 135,000 SFR homes sales per year. Tampa Bay region (including Hillsborough, Pasco, Pinellas, Manatee, Sarasota, and Charlotte counties) has been chosen as ad-hock check on the data quality and graphed along the Florida SFR home sales. It is evident that Tampa Bay region follows overall Florida trend. Please, note the sharp drop in sales from 2019 to 2020 for both Florida and Tampa Bay Region.

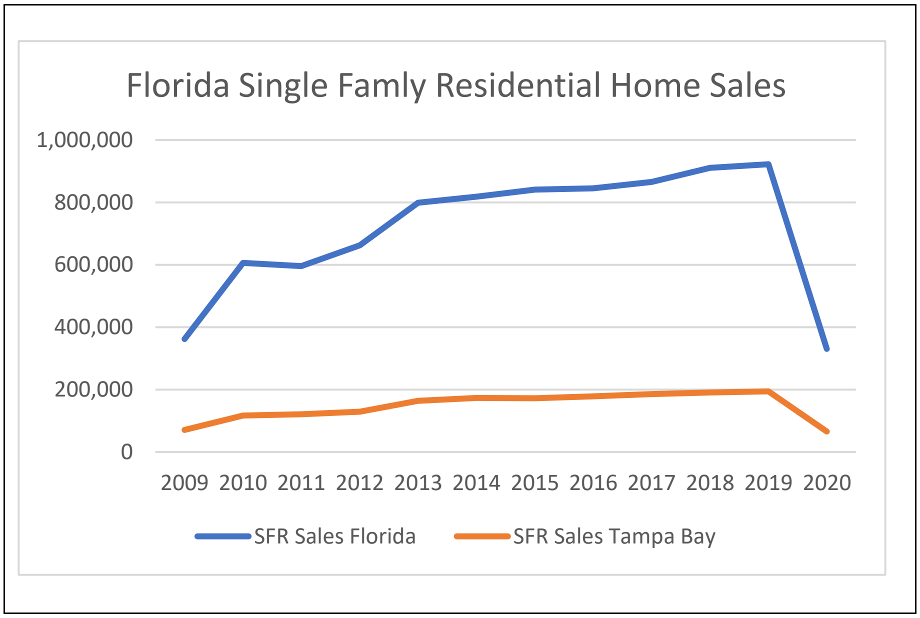


Figure 1 - Florida SFR Home Sales from 2009 to 2020

Zeroing in on SFR Home Same Year Re-sales

FDOR NAL files contain information on SFR home re-sales – each record has *Sale Year/Month One* and *Sale Year /Month Two*. To eliminate the scenario where land parcel sales and new home gets built and sells again in the same year, the data were filtered in the following way: *Sale Year One* not equal to *Actual Year Built*. To determine if the SFR sales with profit or loss – *Sale Price One (last sale)* was subtracted from the *Sale Price Two (previous sale)*. Time between *Sale 1* and *Sale 2* measured in months and was calculated by subtracting *Sale Month One (last sale)* from *Sale Month Two (previous sale).* Detailed descriptions of all fields could be found in the Appendix 1 - 2022 User’s Guide.

NOTE: *The SFR homes bought and sold again, which crossed year boundary, were removed from the study.*

Graphing Florida SFR Home Same Year Re-Sales from 2009 to 2020

The graph below shows counts of Florida SFR home same year re-sales from 2009 to 2020. There were on average – 35,500 SFR homes re-sales per year. Tampa Bay region (Hillsborough, Pasco, Pinellas, Manatee, Sarasota, and Charlotte counties) has been chosen as an ad-hock check on the data quality and graphed along the Florida SFR home re-sales. It is evident that Tampa Bay region follows overall Florida trend. Please, note the sharp drop in re-sales from 2019 to 2020 for both Florida and Tampa Bay Region.

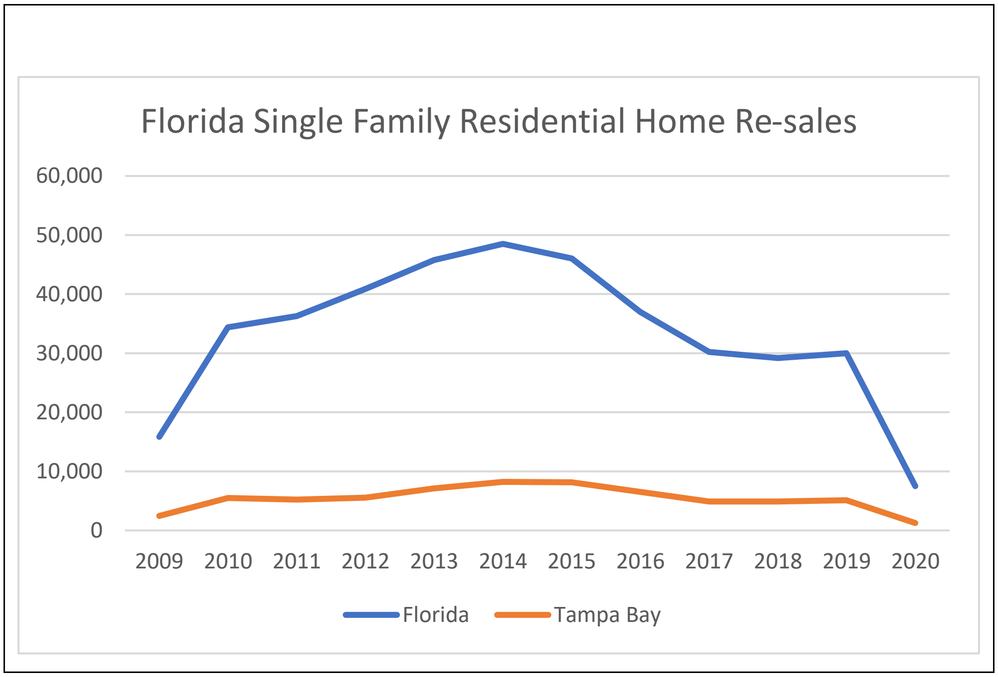
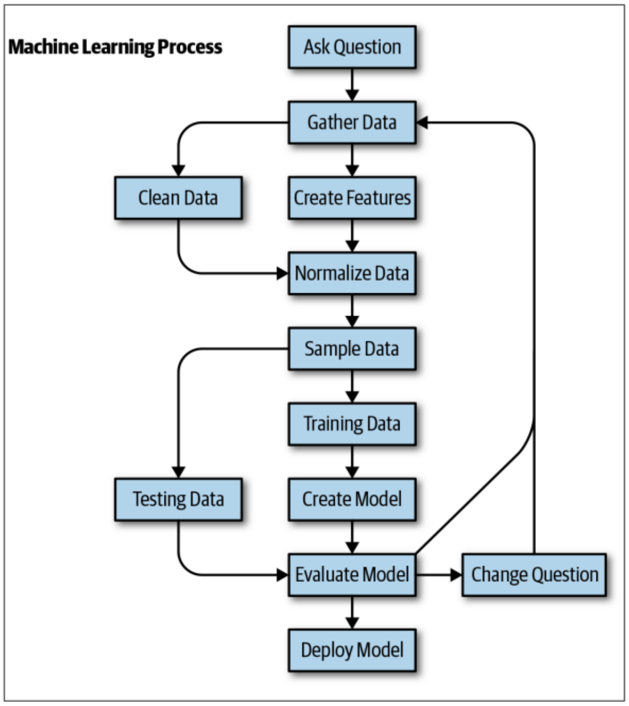


Figure 2 - Florida SFR Home Sales from 2009 to 2020

General Machine Learning Process



Question: Will Single Family Residential Home re-sell for loss or make a profit in the same year?

Data: Florida Department of Revenue Assessment Roll files from 2009 to 2020

The study will be following steps of the General Machine Learning process diagramed on the left in the figure 3.

Logistic Regression, Decision Tree and Random Forest algorithms will be used to create explainable ("White-Box") models.

Figure 3 - Matt Harrison, Machine Learning Pocket Reference Working With Structured Data in Python, 2019

Target Variable – Loss or Profit

Graph below shows counts of outcome of SFR home same year re-sale from 2009 to 2020. Loss was at 89,245 or 22% and Profit – 312,478 or 78%. In cases where *Sale Price One (last sale)* was equal to *Sale Price Two (previous sale),* the outcome was labeled as *Loss.* The rationale was because of the additional expenses associated with the *Cost of Sale,* break-even would still mean money lost.

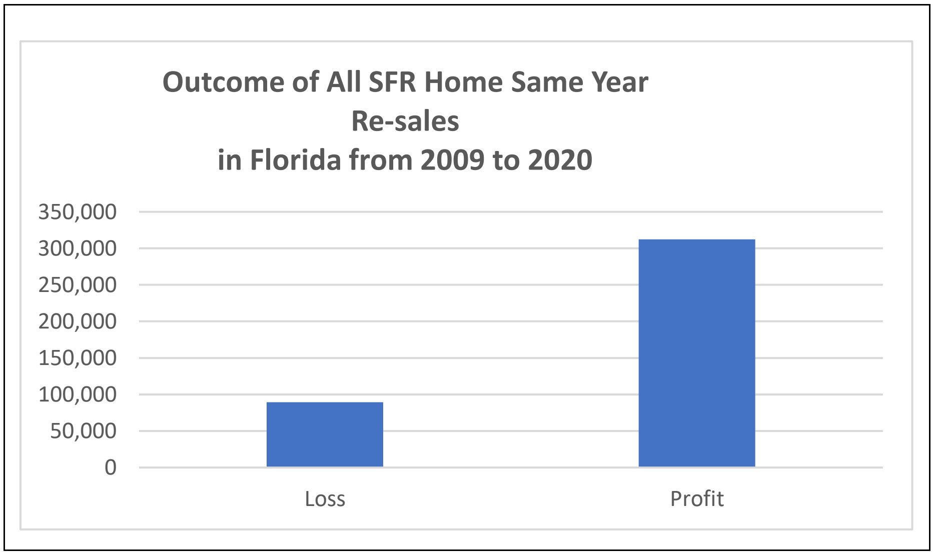


Figure 4 - Outcome of All SFR home same year re-sale in Florida from 2009 to 2020

Target Variable – Loss or Profit (continued)

It is instructive to look at the target variable broken by individual year. Please notice, SFR home re-sale Losses picked in 2013 prior to the highest Profit Year of 2014. The data is uniformly distributed throughout the years with exception for the 2009 and 2020 years.

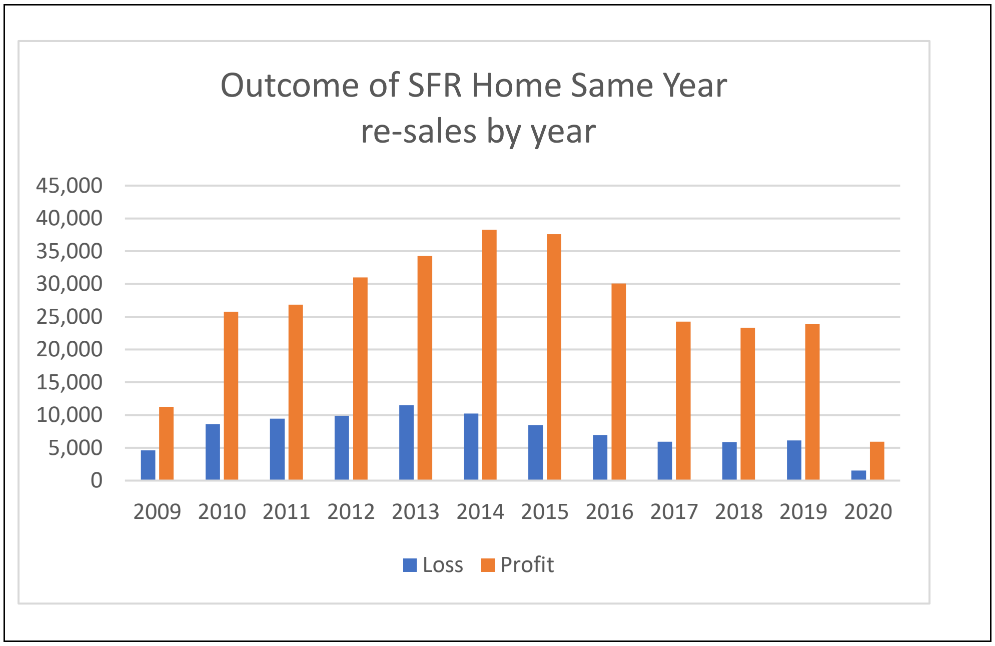


Figure 5 – Outcome of SFR home same year re-sale by year

NOTE: *The SFR homes bought and sold again, which crossed year boundary, were removed from the study.*

Measuring Imbalance of the Data Set

The Imbalance Ratio formula, size of minority class divided by the size of the majority class, was used to measure the extent of the imbalance of the data set. In our case, imbalance ratio was 29% and could be interpreted as for each Loss outcome there are three Profit outcomes. It can be visually seen on the graph above – Figure 5. Upon reviewing Imbalance Ratio, we can conclude - no significant imbalance issue present in the data.

Baseline Accuracy – 78%

Baseline Accuracy has been calculated by dividing counts of Florida SFR home same year resales with outcome of Profit (majority class) by total counts of Loss and Profit outcomes, which equals to 78%. In another words, if we make prediction of Profit 100% of the time, we would be 78% correct. Machine Learning algorithms (Logistic Regression, Decision Tree, Support Vector Machine) will attempt to make predictions with accuracy higher than the baseline of 78%

Predictor Table

It is good practice to attempt pick input variables based on domain knowledge and logical reasoning prior to building models. Following table attempts to estimate positive or negative effect on the Profit or Loss outcome of the SFR home same year re-sale. More detailed descriptions of input variables can be found in the Appendix 1 – 2022 User's Guide

Predictor	Effect	Rationale
LND_SQFOOT	+	Larger land usually sells for a higher price
EFF_YR_BLT	+	Effective Year Built helps homes which are maintained to sell for a higher price
ACT_YR_BLT	-/+	Actual Year Built can positively or negatively effect profit or loss and uncover effects of renovations
TOT_LVG_AREA	+	Larger homes usually sell for a higher price
S_LEGAL	+	Subdivisions within neighborhoods could be more desirable than others
JV	-/+	Just Value could affect investor's decision
IMP_QUAL	+	Higher quality homes usually sell for a higher price
LND_VAL	-/+	Higher land value could affect investor's decision
NO_BULDNG	+	Higher number of buildings usually sell for a higher price
CONST_CLASS	+	Higher class of construction usually sell for a higher price
NCONST_VAL	+	New Construction value added for renovations positive effect re-sale
DEL_VAL	-	Cost of demolition negatively effects profit
SPEC_FEAT_VAL	+	Value of Special Features such as yard items, pool etc. increases the profit
MONTH_DIFF	-	Number of months between Sale 1 and Sale 2 decrease profit
SALE_PRC1	-/+	Sale price of the last sale could help uncover the effect on the profit or loss based on the price point for a particular segment of the SFR homes

Figure 6 – Predictor Table

Feature Engineering

Actual Year and Effective Year Built categorical columns were transformed to numeric columns – Actual and Effective Age. Actual Year Built was subtracted from current year, 2022, for new Actual Age variable. In the similar way, Effective Year Built was transformed into Effective Age column. Actual Age is the number of years from the actual year built and may or may not reflect actual condition of the SFR home. Effective Age column considers maintenance and renovations for the SFR home which were completed over the years.

Data Preparation

Prior to Data Preparation steps, data set's dimensions were 16 columns (including target variable – 0 for Profit and 1 for Loss) and 401,723 rows.

For Logistic Regression, Decision Tree and Random Forest algorithms, inputs were processed in the following way: numerical columns were standardized, and missing values replaced with median for each column; categorical columns were "one hot encoded" - for each unique categorical label new binary (1,0) column were created and missing values were ignored. Models were created with programing language *Python* using *Jupyter Notebooks* and *Scikit-Learn* module.

Processed data set's dimensions were 100,785 columns and 401,723 rows. For each model, processed data was split into Training and Test subsets –70% (100,785 columns, 281,206 rows) and (100,785 columns, 120,517 rows) 30%

Descriptions of Input Variables

Predictor	Data Type	Description
EFF_AGE	Numeric	Effective Age
ACT_AGE	Numeric	Actual Age
LND_SQFOOT	Numeric	Land Square Footage
TOT_LVG_AREA	Numeric	Total Living or Usable Area
S_LEGAL	Categorical	Short Legal Description (Name of the Subdivision)
CONST_CLASS	Categorical	Construction Class
IMP_QUAL	Categorical	Improvement Quality
JV	Numeric	Just Value (Assessment Value)
LND_VAL	Numeric	Land Value
NO_BULDNG	Numeric	Number of Buildings
NCONST_VAL	Numeric	New Construction Value
DEL_VAL	Numeric	Deletion Value
SPEC_FEAT_VAL	Numeric	Special Feature Value
MONTH_DIFF	Numeric	Month Difference between Sale 1 and Sale 2
SALE_PRC1	Numeric	Sale Price – Sale 1
TARGET_VAR	Binary	Profit - 0 or Loss – 1

Figure 7 – Descriptions of Input Variables

MODEL I: Logistic Regression

For the Logistic Regression algorithm’s training data set, baseline accuracy was calculated at 78% For the Training Data set model made predictions with 82% accuracy. For the Test Data set model made predictions with 80% accuracy. One percent difference between Training and Test data sets accuracy measures suggested no model underfitting or overfitting issues.

Classification Matrix for the Test data set

	Predicted Class	
Actual Class	***Profit***	***Loss***
Profit	**90,423**	**3,409**
Loss	**20,693**	**5,992**

Looking at the Classification Matrix above, 3,409 observations model predicted as Loss when they were Profit outcomes. 20,693 model incorrectly classified as Loss outcomes when they were in fact – Profit.

Based on the 80% accuracy and error analysis we can conclude that the Logistic Regression algorithm performed favorably compared with the baseline accuracy of 78% and derived formula can be applied to the new data to make predictions.

Feature Importance

Predictor Variable	Coefficient	Proportion of the Loss Outcomes
S_LEGAL	2.423	0.919
EFF_AGE	0.211	0.553
TOT_LVG_AREA	0.168	0.542
CONST_CLASS	0.074	0.518
NO_BULDNG	0.032	0.508
LND_VAL	0.021	0.505
LND_SQFOOT	-0.008	0.498
SALE_PRC1	-0.017	0.496
IMP_QUAL	-0.023	0.494
JV	-0.027	0.493
SPEC_FEAT_VAL	-0.082	0.479
ACT_AGE	-0.173	0.457
MONTH_DIFF	-0.215	0.446
NCONST_VAL	-0.61	0.352
DEL_VAL	-16.754	0

Proportion of the Loss Outcomes was calculated with the inverse logit of the coefficients.

Upon inspection of the coefficients, we can conclude that location plays the most important role in the determining Loss outcome. S_LEGAL variable was the Short Legal Description of the Subdivision where SFR home located and coefficient was 2.423 with proportion of the Loss outcomes – 0.919

If DEL_VAL (Deletion Value) goes up then there little chance of Loss outcome. (Possible during renovations to combine bedrooms, expand living room etc.)

Figure 8 – MODEL I: Logistic Regression, Feature Importance

MODEL II: Decision Tree

For the Decision Tree algorithm’s training data set, baseline accuracy was calculated at 78% In another words, if we make a prediction of “Profit” for every observation we would be 78% accurate. For the Training Data set model made predictions with 83% accuracy. For the Test Data set model made predictions with 82% accuracy. One percent difference between Training and Test data sets accuracy measures suggested no model underfitting or overfitting issues.

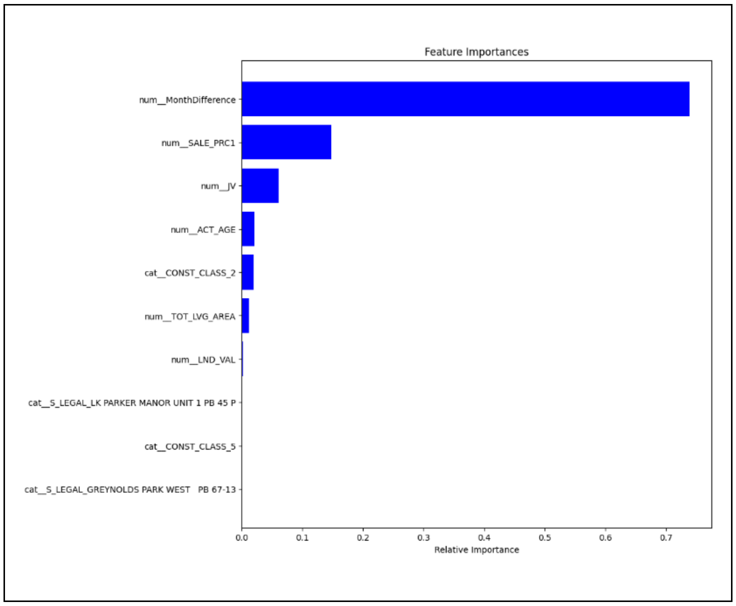
Classification Matrix for the Test data set

	Predicted Class	
Actual Class	***Profit***	***Loss***
Profit	**87,921**	**5,911**
Loss	**15,273**	**11,412**

Looking at the Classification Matrix above, 5,911 observations model predicted as Loss when they were Profit outcomes of the SFR home same year re-sale. 15,273 model incorrectly classified as Loss outcomes when they were in fact – Profit.

Based on the 82% accuracy and error analysis we can conclude that the Decision Tree algorithm performed favorably compared with baseline accuracy of 78% and derived formula can be applied to the new data to make predictions.

Feature Importance



Feature Importance function of *Scikit-Learn* module were applied following training. Chart on the left, shows top 10 variables with importance scores from 0 to 1.

Per *Scikit-Learn* module documentation, higher importance means that there is higher error when the variable removed from the model.
Top 5 importance scores were for the *MONTH_DIFF, SALE_PRC1, JV, ACT_AGE, CONST_CLASS* variables – 0.74, 0.15, 0.06, 0.02, and 0.02

NOTE: *Chart in the Figure 9, derived using random sample of 30% from the processed data set*

Figure 9 – MODEL II: Decision Tree, Feature Importance

MODEL III: Random Forest

For the Random Forest algorithm’s training data set, baseline accuracy was calculated at 78 For the Training Data set model made predictions with 99% accuracy. For the Test Data set model made predictions with 84% accuracy. Fifteen percent difference between Training and Test data sets accuracy measures suggested model overfitting issues. However, the Random Forest made less errors on the test data set compared to Logistic Regression and Decision Tree algorithms and can be considered the best model.

Classification Matrix for the Test data set

	Predicted Class	
Actual Class	***Profit***	***Loss***
Profit	**90,508**	**3,324**
Loss	**15,614**	**11,071**

Looking at the Classification Matrix above, 3,324 observations model predicted as Loss when they were Profit outcomes. 15,614 model incorrectly classified as Loss outcomes when they were in fact – Profit.

Based on the 86% accuracy and error analysis we can conclude that the Random Forest algorithm performed favorably compared with the baseline accuracy of 78% and derived formula can be applied to the new data to make predictions.

Feature Importance

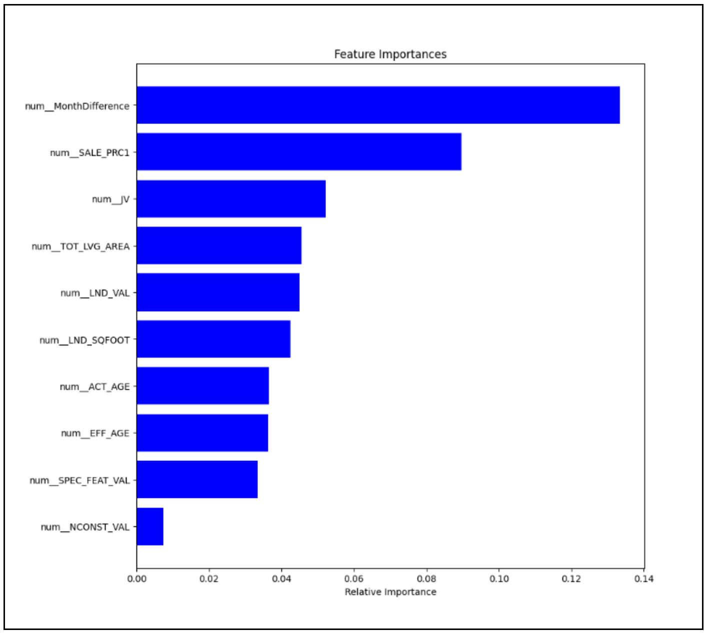


Figure 10 – MODEL III: Random Forest, Feature Importance

Feature Importance function of *Scikit-Learn* module were applied following training. Chart on the left, shows top 10 variables with importance scores from 0 to 1.

Per *Scikit-Learn* module documentation, higher importance means that there is higher error when the variable removed from the model.

Top 5 importance scores were for the *MONTH_DIFF, SALE_PRC1, JV, TOT_LVG_AREA, LND_VAL* variables – 0.133, 0.090, 0.052, 0.045, and 0.042

NOTE: *Chart in the Figure 10, derived using random sample of 30% from the processed data set*

Discussion

The question – will the investment create a loss or profit for the Single-Family Residential home same year re-sale can be answered with 86% accuracy based on the analysis of Florida Department of Revenue's Assessment Roll data from 2009 to 2020. The pure guess would produce 78% accuracy which is only 8% away from the best model's prediction.

Important finding of the research is the discovery of the SFR homes same year re-sale population's Loss or Profit outcomes – 22% and 78% Project and data preparation process combined ten years of data into one data set. The combined data set contains data on every real estate property parcel in Florida from 2009 to 2020. Many questions can be asked and answered of combined data. The research focused on uncovering effects on SFR home same year re-sales outcome.

Three models were created: Logistic Regression, Decision Tree and Random Forest. Logistic Regression model's accuracy on the test data set were 80%, Decision Tree's – 82%, and Random Forest's – 84% All three models were chosen because of their ability to measure each input variable's importance.

Conclusions

Following training feature importance function of *Scikit-Learn* machine learning module were applied to Decision Tree and Random Forest model's derived formulas. Per *Scikit-Learn* documentation, higher importance means that there is higher error when the variable removed from the model. Both models showed high importance scores for MONTH_DIFF variable at 0.74 and 0.13 (MONTH_DIFF variable is the month difference between Sale 1 (last sale) and Sale 2 (previous sale))

Logistic Regression model's coefficients support conventional proclamation: "*Location, Location, Location!",* as it relates to the real estate value, meaning identical homes can increase or decrease in value due to location. S_LEGAL variable's coefficient – 2.42, was the highest, with proportion of the Loss outcomes – 0.92 (S_LEGAL variable is the Short Legal Description of Subdivision)

Lowest coefficient was for DEL_VAL (Deletion Value) variable – (-16.75) with proportion of the Loss outcomes – 0.00; Second lowest coefficient was for NCONST_VAL (New Construction Value) – (-0.61) with proportion of the Loss outcomes – 0.35; Deletion and New Construction Value possibly can be attributed to renovations completed between Sale 1 (last sale) and Sale 2 (previous sale)

Lessons Learned

During research and analysis there were many lessons learned, errors discovered, and many back-and-forth steps taken. Following attempt to enumerate most memorable take aways from the project.

Lesson one: write code and test models on the small sample of the dataset first before applying models on full data set. This lesson was learned when models stopped running over night and operating system asked for more memory, which resulted in delays. In the end, models showed little difference between full population and sample data sets.

Lesson two: write a plan and try to avoid deviating from it. There were temptations to expand original scope of research, creating more models, and getting more data from 2002 to 2009 years. Multiple back-and-forth data preparation and inputs variable selection steps delayed project.

Lesson three: use version control. Version control platform such as Git Hub helps synchronize multiple environments (laptop, work and home PCs) and make changes consistent. Without version control it is easy to end up with multiple versions of Project Report, Data Models and Data sets. Git Hub public repository where INPUTS data sets, Jupyter Notebooks with Python code and final report can be cloned from -https://github.com/Rusonmed/FloridaFlippers_VersionControl.git