GB 565 / FINAL PROJECT

Kaggle Competition: House Prices - Advanced Regression Techniques

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

Group #7

CARSON BATCHELOR
LAINAH MANGWIZA
NOAH PELBERG
MARCELA SALAZAR H.

GB565 / Final Project / Group 7

Kaggle Competition: House Prices - Advanced Regression Techniques

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

I. ESTABLISHING THE BUSINESS CASE

Stepping into the offices of "Cardinal Heritage Realty", Chip Montgomery IV felt anxious. He had been here many times as a kid, but he was not here for fun and games anymore he was here to modernize a 100-year-old family company. Since 1924, “Cardinal Heritage Realty” has been Ames’ trusted realtor, the oldest realtor in Iowa. Chip is just back after graduating from a MSBA at UW-Madison. Now, with his degree in hand he feels the responsibility to revolutionize the family business.

So far, the agency has survived many obstacles; its motto “Our Clients Are Our Family” reflects the agency’s value towards tradition. Chip’s grandfather often tells his grandson: "You know, son, I've been doing this so long, I just see a house and know exactly how much it should sell for." His father relies on more traditional methods to price houses, often pricing through Comparative Market Analysis and his licensed appraisal expertise.

Nevertheless, Chip has been looking at the books and the numbers are not good. The firm’s traditional 5%-6% commission on sales is not the issue, it’s the constant and progressive decline in market share that has led to these negative results. Potential clients are leaning towards national realtor chains like Remax or Zillow. While Chip’s family attribute this to "changing times," Chip sees a different story in the data: their traditional pricing methods, while personal, often result in homes staying on the market longer than necessary, which discourages customers, who decide to look for other options.

After several discussions Chip finally got to present his idea “Heritage Analytics”, a tool that would blend tradition and machine learning algorithms to properly price properties. By utilizing Heritage Analytics Chip’s goal is to obtain a 20% reduction in the average days on market (DOM) for homes in Ames. Currently, homes take an average of 45 days to sell. By pricing properties more accurately, Chip aims to make the agency more attractive for sellers, as their houses would sell 9 days earlier (36 days from listing to sales). If properties are correctly priced and sell rapidly this can lead “Cardinal Heritage Realty” to regain market share in Ames, and who knows even expand to neighboring cities like Gilbert, Nevada, Huxley and Story City.

II. CITY OF AMES

Ames, Iowa is a city with a population of approximately 66,265 people. The city, founded in 1864, is the result of the expansion of the Cedar Rapids & Missouri River Railroad. Through the use of an API Key, Chip could retrieve data for the city of Ames from the American Community Survey 2018-2022 (5 year estimate) published by the US Census bureau. According to the US Census Bureau, the city of Ames has a total of 27,356 residential housing units, of which 2,216 are vacant, which translates to a vacancy rate of 8%. The 92% housing occupancy rate signals a strong housing demand. Renter-Occupied Housing represents 58% of all occupied units approximately. It is relevant to mention that Ames is home to Iowa State University, which enrolls an approximate of 30,000 students. Therefore, housing options accommodate both students and permanent residents. This renter-heavy market represents a promising market opportunity for Cardinal Heritage. Housing units include apartments, houses and townhouses.
Through the ACS data, Chip also collected information about Median Home Value, which stands at $247,500.00, higher than the state average ($124,249.47) providing further evidence that Ames has a vibrant real estate market.

```
#API_KEY = "1696cc44a6ddb2a578cfe6cff13bc17e0ac6ea70"
#2018-2022 ACS 5-year data = base_url = "https://api.census.gov/data/2022/acs/acs5"
```

III. THE REALTY MARKET

The concept of the realty market in the US dates to the colonization era, with land transactions among settlers and Native Americans. From there the real estate brokerage kept developing, until the creation of the National Association of Realtors (NAR) in 1908 in Chicago. The U.S. real estate and brokerage market is projected to be the highest valued globally in 2024, standing at an astonishing $132 trillion.

IV. HERITAGE DATA

After solving some issues with the on-premise server, and several SQL queries, Chip was able to obtain a CSV file that contained 81 columns and 1460 rows of data; this file would serve him as his training data. He also queried a CSV called test.csv that contained 1459 rows and the same amount of columns, this file will serve him to test the models predictions. This was the only data that he could retrieve as some data was lost due to server malfunction. Carinal Heritage had no contingency plan, resulting in lost data.
After this project was complete Chip was definitely looking into Cloud based solutions, like Amazon Web Services, where the data could be safely stored and retrieved without it being such a burden.

V. DATA PREPARATION

A rigorous data preparation process was implemented to ensure the dataset was clean and suitable for modeling. Initially, missing values were analyzed and visualized using a heatmap to identify patterns of absence. The next step involved excluding columns exhibiting more than 100 missing values. The removal of these columns was to ensure their presence did not introduce potential biases or distort the model's predictive reliability.

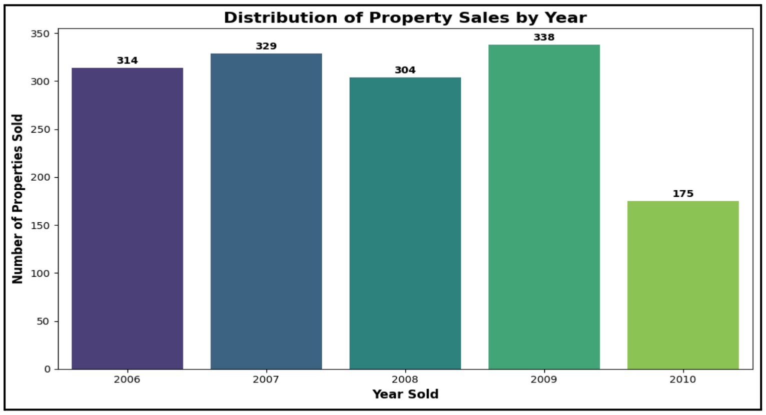
For the remaining missing data, median imputation was employed. This approach was selected due to its ability to ensure the dataset's integrity through maintaining the dataset's central tendency, while minimizing sensitivity to outliers. Numerical variables with missing data were imputed through this method to retain their underlying distribution.

Subsequently, label encoding was applied to encode categorical variables. This transformation converted the data into numerical formats suitable for machine learning algorithms to process and analyze these features effectively. Notable categorical variables, such as *Neighborhood* and *OverallQual* were processed and encoded to preserve the categorical relationship while enabling model compatibility.

Having completed these systematic data preparation steps. The dataset was primed for training, establishing a strong foundation for modeling and accurate predictions.

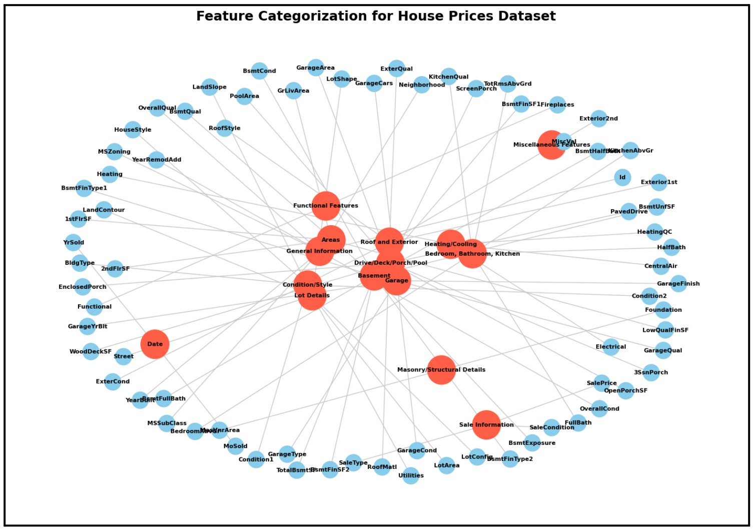
VI. WORKABLE DATA

Before getting his hands "dirty" Chip proceeded to check his cleaned data. He still had 1460 rows. The years the data covered were 2006 to 2010. He obtained this information by looking at the 'YrSold' column. The graph shows the distribution of houses sold for each year. Chip was a little disappointed this is the best data that he could obtain. Nevertheless, he was not going to admit defeat, he planned on implementing a strategy, such as a Market Index Adjustment or Consumer Price Index adjustment to align with 2024 market conditions.



After laying the groundwork towards the goal of achieving predictive accuracy. Chip has a dataframe with seventy four features, each of which provides information related to the properties. These variables range from general property characteristics to very specific property details that might increase/decrease the value of the property.

These seventy-four features are the foundation for feature selection and feature engineering. Chip has the challenge of understanding the features, to do so in an organized manner, he proceeds in a traditional manner, using categories used for Comparative Market Analysis to classify the features. The graph shows the categories (in red) and how each variable was categorized.



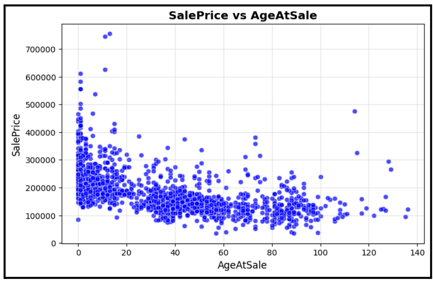
VII. FEATURE ENGINEERING

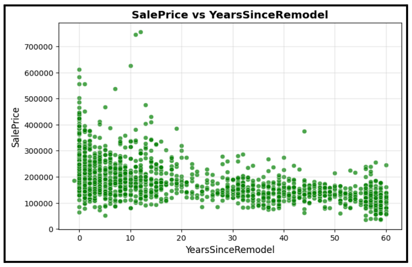
After looking at the variables available to build the house sale prediction model, Chip decides to engineer some features. The features engineered were based on three relevant aspects: Business knowledge, Technical Expertise and Market Relevance.

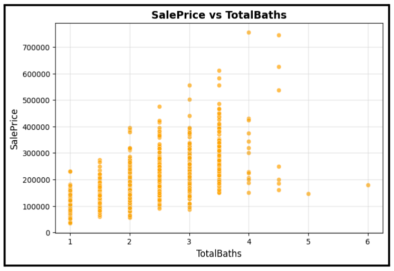
N.	Engineered Feature	Formula	What does this Feature Contribute?		
			Business	Market	Technically
1	Age at Sale	AgeAtSale: (YrSold – YearBuilt)	The age of a property is a key factor for pricing real state. Houses that are newer are more attractive to buyers as they often require less mainttenance or remodeling.	As houses get "older", the likelihood of upgrades increments. Buyers want to pay less for "older" houses.	Age at Sale will become a continous predictor in our model, establishing a clear relationship to price.
2	Years Since Remodeled	YearsSinceRemodel: (YrSold – YearRemodAdd)	A similar concept to Age at Sale, homebuyers perceive remodeled houses as requiring less mainttenance. This makes them more attractive to potential buyers.	The National Association of Realtors highlights recent renovations as a major selling point impacting home valuation.	Year Since Remodeled will become a continous predictor in our model, establishing a clear relationship to price.
3	Total Bathrooms	TotalBaths: FullBath + (HalfBath * 0.5) + BsmtFullBath + (BsmtHalfBath * 0.5)	The total number of bathrooms is a key aspect for potential homeowners to consider. Multiple bathrooms in a property often lead to higher pricing.	Potential buyers see basement bathroom as space that is functional. Including it in our feature engineering is key.	From 4 features we reduced this key data to 1 feature. Reducing the data dimensionality, creating a model that is more interpretable.
4	Total Porch Area	TotalPorchSF: OpenPorchSF + EnclosedPorch+ 3SsnPorch + ScreenPorch	This responds to a market trend where homebuyers are placing higher impoirtance on outdoor activities, according to NAR.	Porch is considered usable are. Homeowners are not interested in type of porch but instead assign vale to the total porch area.	From 4 features we reduced this key data to 1 feature. Reducing the data dimensionality, creating a model that is more interpretable.

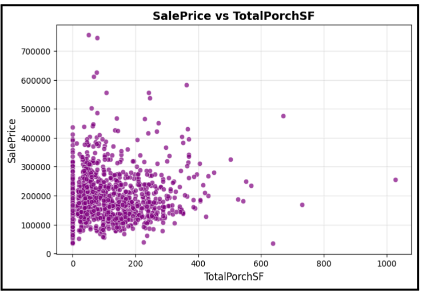
***Source:** NAR Remodeling - https://www.nar.realtor/remodeling
NAR Porch Space - https://www.nar.realtor/blogs/styled-staged-sold/the-reinvented-screened-in-porch-offers-flexibility-for-the-seasons

After defining the engineered features to utilize, Chip proceeds to plot each of the features against SalePrice (the target variable), this is performed to validate feature relevance. A clear trend either positive or negative can validate that the features are relevant for Chip to use in the price prediction model.









- Sale Price vs. Age at Sale: Shows a negative correlation. Houses that are newer show a higher sale price. This feature will likely become a high predictor.

- Sale Price vs. Year Since Remodel: Lower prices are clustered the further we move from zero on the x-axis, meaning the older the remodeling took place.

- Sale Price vs. Total Baths: Shows a positive correlation, properties with more bathrooms tend to sell for higher prices.

- Sale Price vs. TotalPorchSF: Most of the properties in our data set have a total porch area of 0 to 200, data is clustered at this range.

- All engineered features are showing a meaningful relationship when compared to Sale Price. With these insights Chip feels confident to include these features

VIII. MODEL SELECTION

From the previous analysis, Chip is certain he wants to include the engineered features. Now he has to choose which features to include from the 74 additional features. He does using a Random Forest Regression model. The results are R-squared:0.9004 , RMSE: 27636.1783. Chip knows his model can perform better than this initial approach, so he looks at the Top 10 most influential variables in the model. From this the chosen features are: OverallQual, GrLivArea, 2ndFlrSF, TotalBsmtSF, BsmtFinSF1, 1stFlrSF, LotArea, GarageArea, GarageCars, Neighborhood. Chip knows that the selected features affect the property value perception from customers. Potential clients assign value to property/infrastructure quality, location and property amenities. This is further enhanced by research by the National Association of Realtors. The chosen variables are all numerical except for ‘Neighborhood’. From the chosen features we will run a model that contains the top 10 general features from our starting Random Forest Regression and the engineered features. The label for the model is the Sale Price.

IX. RANDOM FOREST MODEL

Chip’s goal is to provide Cardinal Heritage Realty with a tool that provides accurate property predictions, reducing the properties average days on market obtaining higher customer satisfaction that results in continuous market share growth. A Random Forest Regression is a model through which Chip can obtain accuracy, accurate pricing for listings will make buyers and sellers prefer Cardinal Heritage over other property options.

Technically Random Forest Regression provides various advantages:

1. Handles Non-Linear Relationships: This is especially useful as some of the variables, like GrLivArea, after a certain point do not increase proportionally as property size increases.
2. Feature Importance Ranking: Helpful both technically and from a business point of view, as Chip will be able to clearly explain to potential customers what features are influencing property pricing the most.
3. Robust to Outliers: Data might include outliers, Random Forest is not very sensitive to outliers making it a good fit.
4. Decrease Overfitting Risk: Random Forest uses bootstrap aggregation, this reduces the chances of overfitting.

To measure the model’s performance Chip will look at two key metrics:

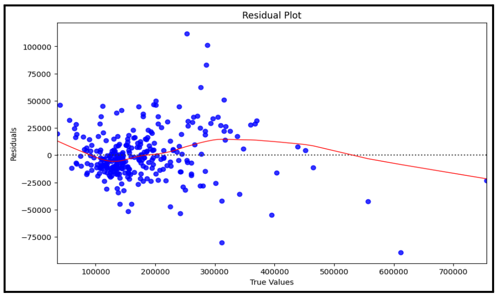
1. Root Mean Squared Error: The lower the RMSE the smaller the possible margin of error when predicting home prices.
2. R-Squared: A high R-Squared signifies that the variance is explained by the model’s feature.
3. Explained Variance Score: Metric that explains how closely the model’s predictions are matching the distribution of real house prices (and not influenced by outliers).

Chip stops for a second and reviews his work so far; he reflects on how his grandfather and fathers achieved success running Heritage; but times do change and a data science approach that mixes tradition with machine learning is best suited. He is sure that by the time his work is finished Heritage will have a model that “sees” houses just as his grandfather did but through a data precision lens.

METRICS	RANDOM FOREST REGRESSION
MSE	26713.91514
R^2	0.906961811
EVS	0.907017822

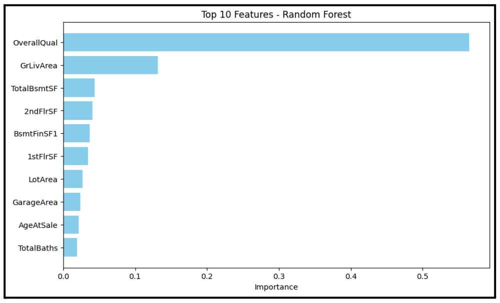
After running the model, the results are:

- A very high R-Squared, 90.7% of the variance is explained by the model.
- MSE is $26,713.91, this error might seem high but when compared to the mean sale price in the train dataset ($180,921.19), the error is 14.77%. This metric is influenced by outliers, nevertheless the percentage is within industry standards.



To further analyze results Chip creates a residual plot from the Random Forest Regression model. Residuals provide important information as they show the difference between the true property values and the predicted property values. Residual visualization allows Chip to assess the validity of the model. The model performs best in the range of $100,000 to $300,000, where most data points are concentrated.

Chip also proceeds to graph the top 10 most important features. These are: OverallQual: 0.564647, GrLivArea: 0.131615, TotalBsmtSF: 0.043705, 2ndFlrSF: 0.040837, BsmtFinSF1: 0.036858, 1stFlrSF: 0.034857, LotArea: 0.026630, GarageArea: 0.023939 and Engineered Features AgeAtSale: 0.021274, TotalBaths: 0.019601.

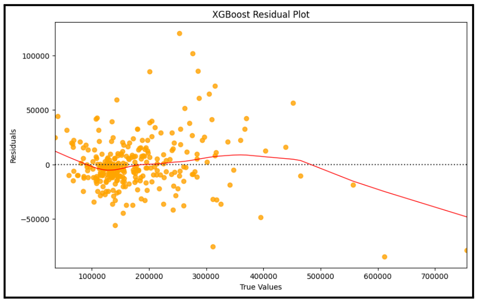


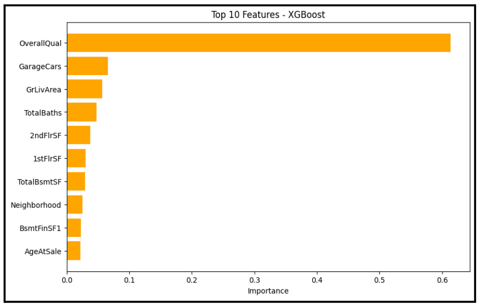
X. XG BOOST

So far Chip is very satisfied with the results from the Random Forest Model, but he remembered that in his Machine Learning class he learned about the strengths of XGBoost. This machine learning algorithm belongs to the family of boosting algorithms, at its core it's the concept of building trees sequentially each of the trees "correcting" the prediction mistakes made by the previous tree (using gradient descent). As with Random Forest Regressor the XG Boost algorithm is also able to handle non-linear relationships, deal with missing values and is robust to outliers. XGBoost also makes some assumptions like: sufficient data quality, independence of observations and some correlation between features. This last assumption is very useful as housing features tend to be naturally correlated.

Running the model he finds very similar results to Random Forest Regressor looking at the same outcome metrics.

METRICS	XGBoost
MSE	25928.78068
R^2	0.912350297
EVS	0.912351143



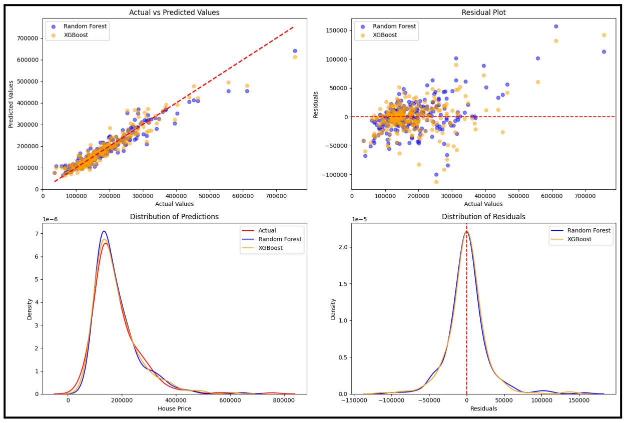


Chip also plots the residuals for XGBoost and looks at the Top 10 features that influence prediction using this algorithm. The feature ranking is different from Random Forest, still the Overall Quality feature ranks the highest.

XI. MODEL COMPARISON

The performance between models only amounts to R-squared = 0.54% and RMSE = $758.00, this is a very small difference. To further decide which model to use when making predictions Chip decided to graph a comparison.

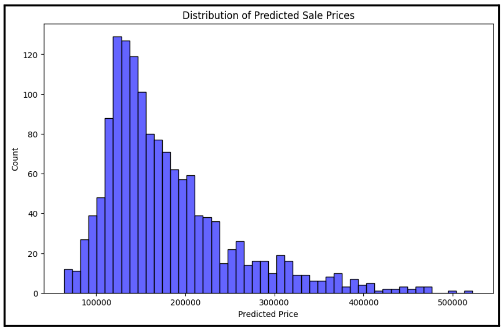
Looking at the graph comparison, we can see both models perform very similarly, but Random Forest predictions are slightly more evenly distributed. Regarding residuals, both models are very similar, with random forest predictions being more symmetric around the zero line.



After comparing models, Chip decides to go forward with his Random Forest model, he bases his decision on: algorithm robustness (the slight more symmetrical distribution suggest predictions that are more stable), simplicity, less risk of overfitting, easier to interpret (and explain to his father and grandfather), fewer hyperparameters to tune.

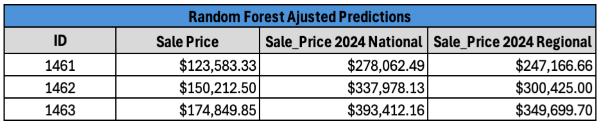
XII. MAKING PREDICTIONS

Now that Chip has decided on a model, he goes ahead and prepares the ‘test’ data for prediction. He performs the same data preparation procedure he used for ‘train’ data, that means: feature engineering, selecting the same features for prediction, making sure the test data frame contains the same number of columns as the train data, encode categorical values and handling missing data. After this process, Chip can finally make predictions on the test data.



XIII. ADJUSTING FOR INFLATION

As Chip realized when looking at the data, some adjustment has to be performed as the dates back to 2010-2016 and he is interested in selling houses in 2024. He does this by using the S&P CoreLogic Case-Shiller Index (Adjustment rate = 2.25) to adjust the prices nationally and goes even further to adjust the prices regionally (using FED data, Iowa Adjustment rate =2), as he knows that Iowa has a lower Index that the national average.



Random Forest Ajusted Predictions			
ID	Sale Price	Sale_Price 2024 National	Sale_Price 2024 Regional
1461	$123,583.33	$278,062.49	$247,166.66
1462	$150,212.50	$337,978.13	$300,425.00
1463	$174,849.85	$393,412.16	$349,699.70

XIV. INSIGHTS

BUSINESS VALUE OF ALGORITHM

Chip knows that his family business has been built on tradition, tradition that has kept the business around for a decade. Because of this, he knows that getting his family on board to make a big change like implementing his model will be no easy task. He decides to show the business value of his model. He knows that drawing more customers to “Cardinal Heritage Realty” means more closings and more revenue. His family knows this as well. Chip feels confident that his family will be convinced of the accuracy and ability of his model to predict housing prices. So, if he can just show them that predicting housing prices more accurately will draw more customers, he feels confident that his family will be on board. Chip set out to survey the prospective home sellers in the city of Ames. During his endeavor, he determined that 90% of prospective home sellers felt that an accurate prediction for what their home would sell for was very important to them. He also determined that speed of sale was a top 3 most important factor in the sale of homes among 80% of prospective sellers. With these two statistics, Chip feels confident that the business value of his algorithm is great enough to convince his family to implement it into their business to help reduce time on market of houses and increase the market share of “Cardinal Heritage Realty.”

SUPERIORITY OF APPROACH TO RELEVANT BENCHMARK METHODS

Chip’s Random Forest Model provides many benefits over traditional appraisal methods like hiring a licensed appraisal or comparative market analysis. Appraisals rely on the subjective opinion of one appraiser, because of this appraisals are subject to inconsistencies and bias. Over time, comparative market analysis came along as a more structured approach to property appraisal; however, this approach still has a heavy reliance on human judgment and does not take into account non-traditional relationships between property features and property price. In addition, both of these traditional models take significant time to come to an appraisal and cost a lot of money. Chip’s random forest model is extremely fast, costs no money and doesn’t rely on human judgement which makes it uniquely scalable. Overall, Chip’s model outperforms traditional appraisal methods by providing an accurate, fast, inexpensive and scalable approach to property appraisal.

XV. CONCLUSION

In conclusion, Chip’s admiration of his family’s business and drive to revolutionize their procedures has played a transformative step in gaining back market share for “Cardinal Heritage Realty.” Chip’s Random Forest Regression model achieves a very high-level of accuracy in providing accurate property pricing predictions. His model provides substantial business value and the reduction of average days on the market by 20% aligns with the priorities of prospective home sellers who value pricing accuracy and sale speed. This positions “Cardinal Heritage Realty” well for regaining their lost market share.

Now that Chip feels confident in his model, he aims to move forward with leveraging this model in his marketing efforts to drive business back to his company. He knows that after implementing his model it will be clear that “Cardinal Heritage Realty” sells homes quicker than their competitors. Once he gains this competitive edge, he knows he can restore his family business beyond where it has ever been.

Sources:

1. City of Ames: https://www.cityofames.org/about-ames/about-ames

2. The Evolution of Real Estate in the United States: A Historical Overview: https://www.markallenrealty.com/blog/the-evolution-of-real-estate-in-the-united-states-a-historical-overview/?

3. Real Estate - United States: https://www.statista.com/outlook/fmo/real-estate/united-states?

4. Zillow Zestimate Tool: https://www.zillow.com/z/zestimate/