# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | House Prices Predicting |
| **Requestor** | Centennial College |
| **Date of Request** | 4th July 2022 |
| **Target Quarter for Delivery** | 26th August 2022 |
| **Epic Link(s)** | Not Applicable |
| **Business Impact** | Make house prices more transparent and help people to properly evaluate house prices |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

This project focuses on enhancing the use of data and information related to housing pricing and makes accurate forecasting and evaluation for all potential house buyers and sellers.

………

**The specific ask:**predicting house prices accurately.

*Clearly articulate the specific task you will be conducting to help achieve the opportunity*

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

House price forecasting is an important topic of real estate. Tryolab is a famous AI pricing company in the current market. Tryolab takes advantage of AI-generated insights and suggestions based on historical data or starts controlled experimentation where data lacks. But the disadvantages are the models Tryolab has applied are not available for retail buyers and sellers and the consultation fees are very high.

Moreover, if asking home buyers to describe their dream house, they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But the dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, it would be easier to produce models for house price predictions.

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

The project attempts to derive useful knowledge from historical data on property markets. Machine learning techniques are applied to analyze historical property transactions to discover useful models for house buyers and sellers. To maintain transparency among customers and also the comparison can be made easily through the models. If the customer finds the price of the house at some given website higher than the price predicted by the model, so he can reject that house.

If we do nothing, all the individual buyers and sellers will not trade houses at proper prices. And property agents can take advantage of asymmetric information to rip off customers. The property markets will be disorganized without some price references.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority, and data availability.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

Make an accurate prediction for house prices in Ames by applying machine learning algorithms.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

This project assumes the models can only be used to price houses in Ames, Iowa.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  |  |
|  | The final prediction accuracy in the testing set should be at least better than random guesses and it would be better if the accuracy is even higher than 65%. And the mean squared error is expected to be low. |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** Regression, Random Forest, Neural Networks

**Methodology:** This project starts with data exploration of the original house dataset such as exploring the distribution of each feature. And I will clean the data such as removing outliers, handling null values, and deleting duplicates. Then I will fit the data into different machine learning models. Next, I will tune the hyperparameters and check the related metrics. Finally, I will analyze the results of different models and compare them to select the best one.

**Output:** The output is house prices.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:** house buyers and sellers

**Observation window:** not applicable

**Inclusions:** data like Electrical system, Central air conditioning, Pool quality

**Exclusions:** macro data like house pricing index

**Data Sources:** https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

**Audience Level:** not applicable

**Variable Selection:** will rank the features and choose not correlated features

**Derived Variables:** will transform categorical data into dummy variables

**Assumptions and data limitations:** current data doesn’t include any macro data such as the current economic situation and CPI

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| Inflation/Deflation will affect prices every year | High |  | we cannot parallelly or directly compare house prices across different years. |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request |  |  |  | 4th July 2022 |
| 2. | Analysis Plan | Come up with detailed planning of the whole project |  | 7 | 15th July 2022 |
| 3. | Data Finalization | Select the final dataset out of three potential topics |  | 7 | 15th July 2022 |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Expore different data features and see their distributions and other related properties |  | 7 | 22nd July 2022 |
| 5. | Modeling | Clean data and fit the data into different machine learning models |  | *14* | 5th August 2022 |
| 6. | Governance | Check related process and documents |  | *14* | 5th August 2022 |
| 7. | Documentation | Analyze the results and write the findings of the project and final report |  | *7* | 12th August 2022 |
| 8. | Peer Feedback |  |  | *7* | 12th August 2022 |
| 9. | Presentation | Prepare PPT slides for presentation |  | 14 | 24th August 2022 |
| 10. | Portfolio |  |  | *14* | 26th August 2022 |