# Analytics Startup Plan

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

| | |
|---|---|
| **Project** | *Car Price Prediction* |
| **Requestor** | *Geely Auto* |
| **Date of Request** | *12th July, 2023* |
| **Target Quarter for Delivery** | *Third Quarter of 2023* |
| **Epic Link(s)** | *Not Applicable* |
| **Business Impact** | *Geely Auto can enhance its pricing strategies to establish competitiveness in the US market by leveraging various prediction models to forecast car price on significant vehicle features with a minimum R-squared value of 0.8 to ensure accurate and reliable predictions.* |

# 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.*

A Chinese automobile company, Geely Auto, wants to set up its vehicle production line in the United States and expand its market share in the US and Europe. This project aims to optimize the price prediction models and forecast corresponding car prices with a minimum R-squared value of 0.8. To increase completeness in the foreign market, it is essential to thoroughly understand the correlation between variables and the manufacturer's suggested retail price. By targeting this goal, high-accuracy prediction models are the prerequisites to assist Geely's pricing and marketing decision-making in the future.

**The specific ask:**

This project needs to recognize and adjust any missing values or typos in the dataset to avoid confusion or unreasonable result at the end. Identifying the correlation between different features and the car price is the next important step toward predicting an accurate vehicle price. Visualizing data patterns and trends also play an essential role in better understanding of the car dataset. By the end of the project, I will compare the accuracy of different prediction models and make necessary adjustments if needed to maximize their efficiency.

# 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?*

Geely Auto is a Chinese automobile manufacturer specializing in affordable models such as sedans, SUVs, and hatchbacks. It has established tremendous prominence in the local car industry and acquired Volvo Cars from Ford Motor in 2010. One of Geely's strengths is combining automotive information with Volvo and gaining access to its innovative ideas, technologies, and global expansion presence. However, Geely's weakness is that they must spend time and money studying the US automobile market to compete with various local competitors. It will be challenging for Geely Auto to fight against different well-known automotive brands in the United States, such as Ford, Toyota, and BMW because they have already built up a stable customer base and positive brand reputation. To penetrate the US market and increase its foreign market share with an appropriate price, a thorough data analytics report is needed for Geely Auto as a reference to adjust its pricing strategy to attract customers overseas.

## 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.*

This project expects up to a minimum R-squared value of 0.8 from one or more prediction models on the car price. Comparing the actual and predicted outcomes to calculate the accuracy rate until an acceptable level has been met. Geely Auto will not be able to penetrate the US car industry with a reasonable price if the price prediction is not accurate and unreliable. Therefore, a high- accuracy prediction model is necessary for the company to determine which pricing strategies are applicable in the foreign market and make use of various prices to expand its market share.

# 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.

Key Questions:

1. Which features have a positive/negative correlation with car price?
2. Which variable has the highest impact on influencing the car price? (Horsepower, Body type etc.)
3. Which prediction models have the highest accuracy rate?

Assumptions:

- Predictor variables and the target variable have a linear relationship. Linear regression or other linear models can be used as a starting point.
- Data authenticity. Ensuring the data source is reliable and useful for Geely Auto to adjust its pricing strategies in the foreign market.
- Predictor variables and the target variable are consistent. Vehicle features and their price do not across various periods or durations. This makes sure the model parameters are constant and avoid time confusion.

Hypothesis:

The vehicle price prediction models in this project are expected to achieve a minimum R-squared value of 0.8 or higher.

## 2.1 Other related questions and Assumptions:

*List any assumptions that may affect the analysis*

Since this small dataset only contains 205 rows, I have made certain assumptions to ensure a smooth data analytics process and mitigate the potential risks.

- This dataset contains high quality data with minimal missing values, typos, and outliers. It means most data accurately represents the relationship between predictor variables and the target variable, minimizing data biases.
- The 200 rows of data consist of random samples that are representative of the population. They encompass a diverse range of observations from the larger population, resulting in reliable results even with a limited number of data points.

## 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.*

This project strives for success in predicting car prices with a minimum R-squared value of 0.8 or higher. Here are some key performance indicators (KPIs) to help keep track of this goal and ensure that every process is on the right track.

KPIs:

- Start with an error tolerance range with ± 8% and lower it down to ± 5% by Q3 2023
- Increase the R-squared value to 0.8 by the end of July 2023
- Test with 90% confidence interval within $5,000 in July and increase to 95% within a range of $3,000 by Q3 2023 (Predicted car price ± $3,000)
- Control the root mean squared error to be less than $5,000 in August
- Set the maximum allowable predicted price difference for mean absolute error to $2,000 in August

## 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:** Linear Regression (1), Decision Trees (2), Random Forest (3), Gradient Boosting (4)

This project will begin by implementing these models with this sequence. Linear regression is a good starting point as it is a straightforward and precise model. Starting with this model provides a baseline interpretation of the relationships between the predictor variables and target variable, showcasing whether they are positively or negatively correlated generates initial insights on

price prediction. As this capstone goes deeper into the data mining, it will gradually increase the complexity of the models to fulfill more advanced tasks. Random forest and gradient boosting will be introduced after linear regression to gain more ideas and insights towards the project's goal and predict car prices on a different perspective.

**Methodology:** This project will perform a data analytics plan that sticks to the key questions in section 2.0 and provides answers by Q3 2023.

After the data cleaning process, we will start by splitting the data into training and validation sets using 0.7/0.3 ratio. Instead of using 0.6/0.4 split ratio that is commonly used by other models, a larger validation set provides an efficient evaluation of the model with limited data size on unseen data. A linear regression model will be built on this train and test split, calculating a linear equation that only contains high priority variables with the corresponding coefficients and the model's accuracy. Random forest and gradient boosting will be applied based on the preprocessed data frame, following by the R2 and RMSE score to check the models' efficiency. At last, we will compare all three models and select the most optimal one with highest score and lowest error. We will also go back and forth to re-test different models to see if there are any room of improvements and make some changes if needed.

**Output:** The project's output will be a unique predicted vehicle price, specifically related to the number of features a car has, such as horsepower, body type, and engine type.

# 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:** Not Applicable

**Observation window:** Not Applicable

**Inclusions:** Horsepower, price, engine type, body type, door, fuel type, fuel system, drive wheel, number of cylinders, peak rpm, highway mpg

**Exclusions:** Car ID, name, aspiration, engine location, wheelbase, car length, width, weight, engine size, bore ratio, stroke, compression, city mpg

**Data Sources:** Multiple market surveys

**Audience Level:** Geely Auto Management

**Variable Selection:** Select the predictor variables based on their feature importance, correlation, and the influence level on car prices

**Derived Variables:** New predicted price

**Assumptions and data limitations:** The dataset does not include the manufacture year of cars, which may result in data biases related to time

# 4.0 Dependencies and Risks

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

| Risks | Likelihood (based on historical data) | Delay (based on historical data) | Impact |
|---|---|---|---|
| *Outliers* | *Medium* | *Low* | *Outliers will create data bias which leads to distorted result or confusion* |
| *Overfitting* | *Medium* | *Low* | *Overfitting leads to overly optimistic outcome and ignores the parts where the model needs to be changed or improved* |
| *Insufficient data* | *Medium* | *Low* | *Less data provides relatively smaller diversity with limited number of data points* |
| *Missing data / null values* | *Low* | *Low* | *Inconsistent data will interfere the models' accuracy and make outcome less reliable* |

# 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 | *Confirmed with advisor that the dataset is tenable, and the learning process is worth studying* | *Dataset found* | *1 week* | *5th July* |
| 2. | Assessment / Triage | *Look for any issues and risks related to this dataset* | *Check data quality* | *1* | *6th July* |

| | | | | | |
|---|---|---|---|---|---|
| 3. | Prioritization | *Data preprocessing* | *Data cleaning* | *2* | *8th July* |
| 4. | Data Exploration & Analysis<br>• Issues with duplicates<br>• Issues with outliers | *Identify data patterns, outliers, missing values, and correlation between the predictor variables and target variable* | *Provide a baseline understanding of this dataset* | *3* | *11th July* |
| 5. | Deploy Training Models<br><br>(Linear Regression, Random Forest, Gradient Boosting) | *Look for the optimal model with highest accuracy and lowest error if possible* | *Select the best model* | *2 weeks* | *25th July* |
| 6. | QA Output | *Mitigate potential risks such as overfitting* | *Increase models' accuracy* | *2* | *27th July* |
| 7. | Model Comparison | *Compare all three models and figure out the underlying reason behind each accuracy and error rate* | *Understand the models' algorithm* | *2* | *29th July* |
| 8. | Improvement | *Be adaptive and flexible to see if there is any room of improvements for different models* | *Further increase models' accuracy* | *2* | *31st July* |
| 9. | Go/No Go | *Double check to see if the outcome is reasonable and reliable* | *Ensure data outcome is accurate and trustworthy* | *1* | *1st Aug* |
| 10. | Storytelling | *Migrate data to Tableau and prepare interactive data visualization for the capstone presentation* | *Data Visualization* | *1 week* | *8th Aug* |
| 11. | Delivery & sign-off | End of capstone project | *Project closure* | *1 week* | *15th Aug* |