MGT 6203 Group Project Proposal Template

TEAM INFORMATION (1 point)

Team #: 85

Team Members:

1. Kenji Hagiwara; 903953592
   - Data Scientist at Grammarly, undergrad in Business Information Systems, previous work in basic ML models and experimentation
2. Shakil Aryal; 903554223
   - Data Scientist/Statistician at EEOC. BS in Economics, Minor in Business
3. Rutvij Rana; 903840445
   - Data Analyst/Strategist in the healthcare industry (previously public sector), BSc in Biological Sciences
4. Akshay Bahl; 903860197
   - Management consultant, undergrad degree in Computer Engineering
5. Samuel Robert Schreer;
   -

OBJECTIVE/PROBLEM (5 points)

Background Information on chosen project topic:

- Irrespective of the role a business assumes—be it as a buyer or a seller of automobiles—predictive analysis remains key to industry practices.
- We will be using the vehicle dataset from Kaggle, to investigate the key variables and their interplay to make accurate predictions of vehicle pricing.
    - Regression models work well with the automotive industry for predictive analysis.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

The relationship between used car sale price and key vehicle attributes is unclear, so the purpose of our analysis is centered around understanding the relationship between the selling price of used automobiles and the independent variables.

Following this, we will use this relationship to predict the selling price of a used automobile based on provided independent variables.

State your Primary Research Question (RQ):

- What is the nature of the relationship between selling price of a used automobile and other independent predictor variables that are part of this dataset? How can we predict the selling price of a user automobile as accurately as possible using the given independent variables?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Compare and contrast a linear regression model against different models in order to find one that best predicts the price of used cars
2. What is the impact to selling price from changing various independent variables?
3. Which variables are the most critical to precisely forecasting the selling price of pre-owned automobiles?
4. Changes to which independent variables have the most significant impact on price?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

- From a business perspective, accurately predicting used vehicle sale prices can be very lucrative.
- For buyers, having access to precise price predictions ensures they make cost-effective purchasing decisions, which can result in substantial savings and increased purchasing confidence.
- For sellers, the ability to set competitive prices can lead to quicker inventory turnover and higher profit margins, contributing to overall revenue growth.
- Streamlining the selling process reduces operational costs and enhances efficiency for dealerships and online marketplaces, resulting in improved operational profitability.
    - Building trust through fair and data-driven transactions can boost customer loyalty and market share, further improving financial outcomes.

- Accurate price predictions also have a profound impact on the automotive industry's financial health and competitiveness.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

Data Description (describe each of your data sources, include screenshots of a few rows of data):

This dataset contains information about used cars.

name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
Maruti Swift	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 20	5
Skoda Rapid	2014	370000	120000	Diesel	Individual	Manual	Second Own	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 15	5
Honda City 2	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700	5
Hyundai i20 S	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at	5
Maruti Swift	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500	5
Hyundai Xcer	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@	5

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Dependent variables

- Selling price: Price at which the car is being sold

Independent Variables

- name: Name of the cars
- year: Year of the car when it was bought
- km_driven: Number of Kilometers the car is driven
- fuel: Fuel type of car (petrol / diesel / CNG / LPG / electric)
- seller_type: Tells if a Seller is Individual or a Dealer
- transmission: Gear transmission of the car (Automatic/Manual)
- Owner: Number of previous owners of the car

New Variables (Interaction terms)

- year and fuel type
- year and seller_type
- year and transmission

Variable Hypothesis

- We assume the most important variable(s) would be the following:
    - year - cars which are newer tend to sell for more
    - km_driven - cars which are less driven indicate less use and therefore a “newer” car

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Data transformations

We can clean up the data to address any null or zero values that would interfere with our results. Additionally, we can create a panel plots of each of the independent variables against selling price to determine if there are any non-linear relationships. In the presence of any nonlinear relationships, transform the corresponding X variable to achieve linearity. We will also use Cook's distance to test for any outliers and depending on the size of the data, remove or impute values to replace outliers. We can also create various interaction terms to better understand whether the relationship between the target and the independent variable changes depending on the value of another independent variable

Approach and models

We propose a multivariate,linear regression to predict selling price. We will compare performance amongst linear-linear, log-log, and linear-log regression. Under a log-linear model the rates change at a constant percent per year (i.e. a fixed annual percent change - APC), while for a linear model the rates change at a constant fixed amount per year. However, when those variables are normal or close to normal, it’s preferred to use a simple linear model.

These are the models we are considering and why we’re choosing the following:

1. Full linear model
    - This serves as the baseline model with all independent variables
2. Log-log regression model
    - Interpretation of the coefficients becomes ubiquitous, in terms of percent. With a log-log transformation, a 1 percent change in X indicates a 1 percent change in the independent variable, holding all other things constant
3. Linear-Log regression model
    - Depending on the distribution of the data, a linear-log model may provide the most value. Traditionally this type of model is used when dependent and independent variables have lognormal distributions.

Train, optimize & compare models

We will be using a Monte Carlo K-fold cross-validation to train, optimize and validate the models to account for important data points present across the input data. We will start with a random split of the data into a training & test data set, for example an 80/20 split. Then fit the models on the train set and predict against the test set to calculate MSPE. We will also record the adjusted r squared for each model for each K as well. We will be using adjusted R Squared or similar metrics to assess fit such as AIC where applicable, to see how much of the variability in selling price is explained by the independent variables. The reason we use adjusted R-squared, rather than R-squared, is so we only consider independent variables which actually have an effect on the performance of the model. After all runs of the K have been run, we will average those MSPE to get an average RMSE for each of the models and compare against each other to see which model had the lowest prediction error. Then we will determine if our hypotheses were also correct.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

Our hypotheses are:

1. Cars with lower age variable value and distance driven are more likely to have a higher selling price
2. There will be certain interactions observed between independent variables such as fuel type and age of vehicle as the perception of value for two cars with the same age is different depending on their fuel type (gas, diesel, EV)
3. Cars sold by individual owners are more likely to be used by only one owner and hence may be better maintained as compared to the dealer-owned cars that may have switched ownership multiple times. This leads to the “seller_type” and “owner” independent variables having a significant effect on the selling price of the car
4. With fewer moving parts, a manual shift has the advantage of being easy to maintain and hence, people perceive that from among two cars of the same make and year, a manual transmission vehicle will have a higher value and hence a higher selling price
5. Cars sold from dealers are more expensive than cars sold by an individual.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

- Areas of a business that can be impacted by the results of this analysis are:
    - Pricing Strategy
    - Inventory Management
    - Supply Chain

- Manufacturing and distribution
- Marketing and Advertising budget

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Week	Date	Progress
1	10/1	Project proposal completed
1.5	10/4 @ 8PM CT	Review project proposal before submission
2	10/7	Division of work distributed & begin modeling work
3	10/14	Modeling work
4	10/21	Model completed
5	10/28	Optimize model
6	11/4	Validate hypotheses
7	11/19	Report complete

Appendix (any preliminary figures or charts that you would like to include):

Model structure

1) Import data
2) Data visualizations
    a) Check relationships & distributions
        i) Panel plots to check the type of relationships
        ii) Correlation heatmaps between selling price and independent variables

3) Transform data
    a) Create our log data sets as needed
    b) Clean up nulls as needed
    c) Create the adequate dummy variables and interaction terms
4) Run Model using k-folds
    a) Create test/train split randomly in each k run
    b) Variable selection
    c) Fit the model, check adjusted r squared, calculate prediction error
5) Compare model performance
    a) Calculated average MSE across the K fold CV
    b) Determine which model had the best fit and best prediction performance
6) Validate each hypothesis based on model performance
    a) Call out omitted variable bias