**Used Vehicle Buying:**

Improving the Buyer and Seller experience through inference and prediction models

**Group Title:** Group1

**Group Members (In Alphabetical Order):**

Ralph Parlin | Patrick Prioletti | Brian Schramke | Kobi Wiseman

**Class:** IST 718 | Professor Willard Williamson

**Table of Contents**

[**1. ABSTRACT** 2](#_Toc54266802)

[**2. DATA** 3](#_Toc54266803)

[2.1 Source and Collection 3](#_Toc54266804)

[2.2 Data Cleaning and Treatment for Missing Values 3](#_Toc54266805)

[2.3 Data Organization 3](#_Toc54266806)

[2.4 Exploratory Data Analysis 3](#_Toc54266807)

[*2.4.1 Summary Statistics:* 3](#_Toc54266808)

[*2.4.2 Data Visualizations and Transformations:* 3](#_Toc54266809)

[*2.4.3 Correlation Analysis:* 3](#_Toc54266810)

[**3. METHOD** 3](#_Toc54266811)

[**4. MODELING** 4](#_Toc54266812)

[4.1 Model 1 4](#_Toc54266813)

[*4.1.1 Model 1 Data Transformations* 4](#_Toc54266814)

[*4.1.2 Model 1 Evaluation* 4](#_Toc54266815)

[*4.1.3 Model 1 Interpretation* 4](#_Toc54266816)

[*4.1.4 Model 1 Inference and Key Findings* 4](#_Toc54266817)

[4.2 Model 2 4](#_Toc54266818)

[*4.2.1 Model 2 Data Transformations* 4](#_Toc54266819)

[*4.2.2 Model 2 Evaluation* 4](#_Toc54266820)

[*4.2.3 Model2 Interpretation* 4](#_Toc54266821)

[*4.2.4 Model 2 Inference and Key Findings* 4](#_Toc54266822)

[**5. CONCLUDING REMARKS** 4](#_Toc54266823)

[**6. ANNEXES** 6](#_Toc54266824)

[**ANNEX 1: Data Dictionary** 6](#_Toc54266825)

[**ANNEX 2: xyxyxy** 6](#_Toc54266826)

[**ANNEX 3 : Python & Pyspark packages used in the analysis** 6](#_Toc54266827)

[**ANNEX 5: Supporting Visualizations** 6](#_Toc54266828)

[**ANNEX 6: Scatterplot Analysis** 6](#_Toc54266829)

[**7. ENDNOTES** 6](#_Toc54266830)

# **1. ABSTRACT**

Bisuabiusabc[i

* Provide a high-level overview of your project
* Provide a list of the specific predictions you made
* Provide a list of the specific inferences you made. In this context, inference is using parameters of a trained model to gain insight about the problems you’re trying to solve.
* Provide a brief conclusion summary outlining the success / failure of your predictions and inference
* Provide a list of other goals your project may have.

The abstract should be no more than 1 to 2 pages.

# **2. DATA**

## 2.1 Source and Collection

We identified a database that contains

A complete Data Dictionary for these tables can be seen in ANNEX 2.

## 2.2 Data Cleaning and Treatment for Missing Values

The data cleaning phase of this analysis revealed the dataset had …..

## 2.3 Data Organization

Before moving to our exploratory data phase, we partitioned the cleaned dataset into three distinct datasets…..

## 2.4 Exploratory Data Analysis

The exploratory data phase consisted of three distinct activities. The first was to provide summary statistics for each of the features in the dataset, the second activity was to create visualizations for both numeric and categorical features, and lastly was to examine the scatterplots and correlations between each feature. Each of these actives helped us better understand the features, their relationships, and provided us some initiation as we prepared to move into the modeling phase.

### *2.4.1 Summary Statistics:*

### *2.4.2 Data Visualizations and Transformations:*

### *2.4.3 Correlation Analysis:*

*Additional Scatterplot Analysis can be found in ANNEX 6.*

* An expanded dataset description from what you provided in the project proposal.
* If you collected the data by scraping a web site, describe that here.
* Describe the results of your data exploration.
* Provide some data exploration visualizations.
* Tell me something interesting you learned about the data.

Provide brief details about what you had to do to get the data ready for training your models. For example, did the data have a lot of NA values. Did you end up dropping columns of data. Did you do feature engineering. Did you use dimensionality reduction.

# **3. METHOD**

Provide a brief description of the methodology used to achieve the goals of your project. For example, if you are performing sentiment analysis on tweet data, describe the high-level methodology of your data science work flow to go from post data wrangled text from the previous section to trained models. Block diagrams work well here. Be brief with words and liberal with diagrams and figures. Some brief text along with a block diagram or 2 should be fine.

# **4. MODELING**

* For each model:
  + Describe what the model is predicting.
  + Indicate the model type: Random Forest, Linear Regression, etc.
  + Briefly describe any special data transformations required for this model. Example: Performed PCA transformation and used the first 20 components which described 80% of the data variance.
  + Briefly describe how you scored the models (Example, MSE, AUC, F1, etc). Describe any special scoring goals and how you achieved the goals. Example, you were trying to minimize false negatives.
  + Briefly describe any inference you performed using the model
* A good technique for this section is to put each model in its own sub section. Briefly describe the items above in words. Follow up with one or more tables that summarize the prediction and inference results. The goal is to provide brief text and summarize your results in easy to read tables.

## 4.1 Model 1

Model 1 was designed to help answer our first research question: *“How do ….. Model 1 is a \_\_\_\_ model.*

### *4.1.1 Model 1 Data Transformations*

### *4.1.2 Model 1 Evaluation*

### *4.1.3 Model 1 Interpretation*

### *4.1.4 Model 1 Inference and Key Findings*

## 4.2 Model 2

Model 2 was designed to help answer our second research question: *“How do …. Model 2 is a \_\_\_\_ model.*

### *4.2.1 Model 2 Data Transformations*

### *4.2.2 Model 2 Evaluation*

### *4.2.3 Model2 Interpretation*

### *4.2.4 Model 2 Inference and Key Findings*

# **5. CONCLUDING REMARKS**

asudfbaiusfbpiuabfiuab

Describe the results of the project by expanding upon the brief conclusion results outlined in the abstract. Summarize model comparison results. Summarize inference results. Provide results for every prediction and inference item presented in the abstract. A good methodology for this section is to be brief with words and liberal with summary tables.

# **6. ANNEXES**

## **ANNEX 1: Data Dictionary**

## **ANNEX 2: xyxyxy**

## **ANNEX 3 : Python & Pyspark packages used in the analysis**

**Python Packages:**

skLearn

matplotlib

pandas

numpy

imblearn

seaborn

## **ANNEX 5: Supporting Visualizations**

## **ANNEX 6: Scatterplot Analysis**

# **7. ENDNOTES**