Final Project Proposal – IST 652

**ANALYZING CRAIGSLIST USED CARS INVENTORY IN THE U.S.**

DATA SET USED

Craigslist is a leading global market for used vehicles. In fact, it is the leading online portal for the sale of used vehicles in the US. The Kaggle Used Cars dataset was built by Austin Reese as a scrapper for the all the vehicles on craigslist website for sale in the US. According to Kaggle, “it contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories. For ML projects, consider feature engineering on location columns such as long/lat.”

Data Source: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

TOPIC OF INVESTIGATION

**What is the value decay of different vehicle models across the U.S?**

This topic of investigation will be used to address the following:

1. **Exploratory Analysis**

This question will explore how vehicles retain their value over time. It will compare manufacturer groups such as Toyota and Dodge while also looking at differences within groups for example Family Vans like Toyota Sienna, and Dodge Caravan.

Additionally, it will explore value decay along state lines for similar vehicle brands. For example, in which state does Toyota vehicles retain their values the most over time? Where do they lose their value the most over time?

1. **Predictive Modeling**

This question will lead to an algorithm that predicts value decay of any vehicle given the state where it is being sold at, it’s perceived condition over time (good, bad, scrap), as well as the number of years projected into the future. Conversely, it will model the pricing behavior of vehicles in the inventory across the United States in order to predict the price of such a vehicle given its age, manufacturer and the state where it is being sold. It will also suggest the best price point to sell a car given its make, model, state where it is being sold and condition.

1. **Text Mining / Sentiment analysis / Natural Language Processing**

The dataset used in this analysis contains vehicle descriptions which is a mix of both structured and unstructured data. These reviews will be used to characterize the vehicles. For example, what are the commonly used words in describing Toyota vehicles? Given a set of words, what is the probability that the vehicle being described is a Toyota or a Ford. This will be achieved using Frequency Distribution of word tokens as well as performing mutual information scores for bigrams and trigrams.

PROJECT SCOPE

The completed analysis will include both tabular reports as well as graphical and map data that speaks to the material under investigation. For example, while exploring vehicle price decay across states, a tabular chart with states, vehicles and price decay over time will be included, along with a map of all the states in the US with varying shades and sizes of price bubbles that depict a vehicle’s price over time.

This project will handle more than 1GB of data, containing both structured, unstructured and semi structured records. It will employ both exploratory and predictive approaches while using quantitative reasoning such as ANOVA. Furthermore, all quantitative reasoning will use Bayesian methods instead of Frequentist approaches.