Machine Learning Canvas

	PREDICTIONS	OBJECTIVES	DATA
	End-user Who will use the predictive system / who will be affected by it?	**Value proposition** What are we trying to do for the system's users? (e.g. spend less time on X, increase Y...)	**Data sources** Where do/can we get data from? (internal database, 3rd party API, etc.)
ƎRO	This predictive system would be used by both car sellers, to price their vehicles competitively, and car buyers, to avoid overpaying for a car.	Help users price used cars more accurately (avoid underpricing or overpricing) to facilitate faster sales and better returns for sellers, and avoid overpaying for buyers.	Public dataset on Kaggle: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data

PREDICTIONS

Problem

Question to predict answers to (on behalf of user)

What is the fair market value of a specific used car?

Input (i.e. question "parameter")

Car attributes such as Make, Model, Year, Mileage, Cor

Possible outputs (i.e. "answers")

Predicted price

Type of problem (e.g. classification, regression, recommendation...)

Regression

Baseline: simple, alternative way of making predictions (e.g. manual rules)

```
Find the most similar entry and return
the price
```

E SF

OBJECTIVES

Performance evaluation

Domain-specific / bottom-line metrics for monitoring performance in production

```
- Average Listing Duration
- Seller Profit Margin
- Buyer Satisfaction
```

Prediction accuracy metrics (e.g. MSE if regression; % accuracy, #FP for classification)

```
- RMSE
- R^2
```

Offline performance evaluation method (e.g. cross-validation or simple training/test split)

```
K-Fold Cross-Validation
```

DATA

Data preparation

How do we get training data (inputs, and outputs if supervised learning)? How many data points?

```
Take the dataset described above, take 20%
sample, and then split this sample into 80%
train and 20% test datasets.
```

Input features (extracted from data sources). If too many, list types of features and mention key ones.

```
- id
- url
- region
- region_url
- price
- year
- manufacturer
- model
- condition
- cylinders
- fuel
- odometer
- title_status
- transmission
- VIN
- drive
- size
- type
- paint_color
- image_url
- description
- county
- state
- lat (latitude)
- long (longitude)
- posting_date
```

PREDICTIONS OBJECTIVES DATA

Using predictions

When do we make predictions and how many?

Predictions would be made whenever a user enters information about a specific used car.

What is the time constraint for making those predictions?

Predictions should be made within seconds after the query is made

How do we use predictions and confidence values?

GRAT

The predicted fair market value would be displayed prominently for users.

Learning models

When do we create/update models? With which data / how much?

First model will created during this project, then the model will be updated as soon as drift happens, i.e. the metrics on unseen data drops significantly. For new model, we will incorporate both current data and unseen data.

What is the time constraint for creating a model?

Semester, i.e. until the end of July

Criteria for deploying model (e.g. minimum performance value — absolute, relative to baseline or to previous model)

- The model should achieve an RMSE of less than $1,000 on the car price predictions.
- The models hould achieve an R^2 of greater than 0.85 on the car price predictions.

Reset Form

Machine Learning Canvas v0.1