Machine Learning Canvas

	PREDICTIONS	OBJECTIVES	DATA
GRO	**End-user** Who will use the predictive system / who will be affected by it? The predictive system will be used by the taxi company to set fair and dynamic pricing. It will affect both the taxi company, by optimizing revenue, and customers, by ensuring reasonable fares that enhance satisfaction.	**Value proposition** What are we trying to do for the system's users? (e.g. spend less time on X, increase Y...) - Increase customer satisfication - Increase revenue of the company	**Data sources** Where do/can we get data from? (internal database, 3rd party API, etc.) - Third-party data from Uber and Lyft databases (historical trip records, fare information). - Third-party APIs for weather history to enhance the model's accuracy.
E SF	**Problem** Question to predict answers to (on behalf of user) How much the ride fom A to B will cost? Input (i.e. question "parameter") Start and end locations, time, weather Possible outputs (i.e. "answers") Taxi ride price Type of problem (e.g. classification, regression, recommendation...) regression Baseline: simple, alternative way of making predictions (e.g. manual rules) Use a fixed rate per mile/kilometer with additional charges based on time of day or weather condition.	**Performance evaluation** Domain-specific / bottom-line metrics for monitoring performance in production - Revenue - Average fare per ride - Customer satisfaction Prediction accuracy metrics (e.g. MSE if regression; % accuracy, #FP for classification) - MSE - MAE - R2 - RMSE - MAPE Offline performance evaluation method (e.g. cross-validation or simple training/test split) - Cross-validation - Training/test split	**Data preparation** How do we get training data (inputs, and outputs if supervised learning)? How many data points? We take that data obtained from Uber and Lyft rides for certain period of time. Then we split it into train and test data. There are as many data points as number of rides for a certain period (138614 for now) Input features (extracted from data sources). If too many, list types of features and mention key ones. We have 57 input features. They include: - numerical (temperature, distance, windSpeed...) - categorical (cab type, timezone...) - text (source, destination, short_summary ...) - time (timestamp, datetime...)
GRAT	**Using predictions** When do we make predictions and how many? Predictions are made in real-time whenever a customer requests a taxi ride. The number of predictions corresponds to the number of ride requests received. What is the time constraint for making those predictions? Five seconds How do we use predictions and confidence values? We use predictions to show how much a ride will cost for users. We are using confidence values to train the model and see how accurate are the predictions.	**Learning models** When do we create/update models? With which data / how much? We update models periodically, relying on data from Uber and Lyft. The first model is created when we want to start our business in a new city. What is the time constraint for creating a model? End of July Criteria for deploying model (e.g. minimum performance value — absolute, relative to baseline or to previous model) - The new model should demonstrate improved performance metrics compared to the previous version - The model should be stable and robust, showing consistent performance across different subsets of data (e.g., through cross-validation)	

Reset Form

Machine Learning Canvas v0.1