DECIDE: A Decentralized Network for Predictive Models

Tim Ogilvie
me@timogilvie.com

November 3, 2017

Abstract

DECIDE is a decentralized network for building better predictive models. The network enables ad-hoc teams of businesses, data scientists and data owners to build and license predictive models without needing to trust each other or a centralized authority. Teams share revenue using smart contracts and offer incentives for continuous improvement. We propose a new token DecisionCoin (DC) to reward participants and incentivize the ecosystem.

1 Introduction

Predictive models are used in virtually every area of the economy and usage continues to grow. Areas that use models extensively include:

- Banks estimate the probability a borrower will default on a loan
- Radiologists want to identify a tumor in a lung CT-scan
- Airport security decide if a passenger deserves additional screening
- Subscription services predict which customers may churn
- Web sites forecast web traffic to accurately assess demand
- Realtors estimate home sale prices
- Marketers estimate the likelihood a user will click on an ad

Building and monetizing predictive models requires three parties: a business with a problem to solve, a data scientist to build the model and data to train the model. DECIDE offers a better opportunity for all three parties, as described in more detail below.

1.1 Businesses: When Algorithms Compete, You Win

Lots of commercial solutions are available but they suffer from similar problems: it's difficult to know which solution has the most effective model. Bad software can have great salespeople, and vice-versa. Similarly, testing a new solution takes time, effort and expertise

DECIDE runs an ongoing competition to determine the best model for a given problem. The network establishes consensus on the best performing model and makes it easy for the business to execute the model.

Data science contests are a proven model for developing predictive models. There are a wide variety of strategies that can be applied to a given data set and it's hard to know in advance what will work best. A large community of data scientists participates in attempting to find the best solution. Kaggle hosts data competitions that regularly attract more than 1,000 teams and individuals. Participants include some of the world best-known data science researchers.

DECIDE will support a myriad of predictive models with strong incentives for continued improvement. Commercial success will strengthen those incentives, creating a virtuous cycle for the businesses that depend on great models.

1.2 Data Scientists: Great Pay for Performance

Monetizing a predictive model can be very lucrative. But while data scientists are key to getting great models, the benefits almost entirely accrue to the business that owns the model. The bounties payable in Kaggle competitions are orders of magnitude lower than their commercial value.

DECIDE turns data scientists into owners. They receive a recurring revenue stream from the models they create. They can also improve existing models for a share in existing revenue streams. This creates natural competition for talented data scientists and directs their energies to the most promising opportunities.

1.3 Data Owners: Convert Data into Recurring Revenue

Training data is the lifeblood of predictive models. Without great training data, you can't build a model. DECIDE offers companies a complete solution to convert their data assets into a recurring revenue stream. Many companies collect data as a byproduct of their core business that can be used as training data for models. But it's tough to know what your data is worth, hard to connect with buyers, and requires trust.

The DECIDE network ensures that data providers get paid full value for their data. It also provides encryption protection for their data. This allows data scientists to build predictive models without sharing highly sensitive data.

DECIDE will support emerging standards like the Ocean Protocol for data owners.

2 Trustless Teamwork

DECIDE provides a framework that allows data scientists, data owners, and businesses to collaborate without needing to trust each other. Businesses or speculators can stake funds towards the creation of specified models, and can offer bounty payments and/or a share of future revenue to attract data owners and data scientists. Smart contracts appropriately allocate funds.

Smart contracts make it easy for teams to work together to build a model without needing to trust each other. The DECIDE network allows data scientists, data owners and businesses to collaborate and share proceeds fairly.

Teams form to solve a specific data problem, that specifies how the model may be licensed be used, how it will be priced, and how much revenue is payable to each member. The team may also choose to provide bounties that are payable to the data scientist with the best performing model on a specified data, or data owners that provide training data for the model.

2.2 Revenue Distribution

When a model is executed, the customer pays in DC for that execution. Those proceeds are distributed to all owners according to their proportional ownership. Note that bounties are handled slightly differently, as described below.

2.3 Bounty Payments

While some data scientists and data owners will want to participate in the ongoing revenue from their models, others may be interested in an immediate bounty payment. Businesses and speculators can create Bounties to purchase Models or Data Sets that meet their qualifications.

Decisions must have a minimum bounty or customer commitment committed before being exposed to the data science community. This will ensure the business problem is high-value enough to attract the attention of talented data scientists.

Bounties act as a loan against future revenue that would have been owed to the recipient of the bounty. The sponsor who paid the bounty will receive the share of revenue due to the data scientist until it has been repaid.

2.4 Incentives for Continued Improvement

Each team can provide incentives for continued improvement by allocating shares that will be awarded to a model that demonstrates better performance. These incentives will ensure that heavily used models will have lots of data scientists

For example, if a model is developed that improves performance by 5%, the owners of the DAO may elect to issue an additional 2.5% of shares to the developer of that model.

2.5 Preventing Over-fitting

DECIDE provides several mechanisms to punish models that have been over-fit to the training data. These models will perform well against the test data set, but won't continue to have the same performance.

The primary mechanism is a test data set that is only used to score after a model has been submitted. This identifies models over-fit to the training data and unsuited for broader usage.

As a secondary mechanism, pre-payments can be escrowed until a certain amount of time has passed or model execution has been performed a certain number of times.

2.6 Model Ownership & Governance

Model ownership is shared among the owners of the DAO. Each owner has a number of shares determining both share of revenue and voting power in any governance proposals. Any owner can initiate a new proposals by staking a small amount of DC and describing their proposed change. Proposals can be any of the following actions:

- Change the licensing terms

- Approve or disapprove a potential customer
- Change the price of model execution
- Change the value assigned to future improvement
- A tender offer to purchase shares at a specified price

Any member can vote in support or against the proposal. After a specified amount of time and a certain number of members has voted, the proposal can be executed. In the event of a failed proposal, the DC that was initially staked is burned. This helps to prevent spam. [N.B. Need to make sure this balances spam prevention with a desire for dialogue/conflicting opinions. Look at similar approaches]

2.6 Model Lifecycle – An Example

The following sequence (entirely fictional and presented for illustration only) illustrates how DECIDE network participants can interact:

1. Alice works for a hedge fund that trades oil stocks. She wants a predictive model to forecast oil prices. She provides a public data set and allots 25% of future revenue to the data scientist that builds the best model. She stakes 2000 DC towards an upfront guarantee.
2. Carol, Dan, Eliot and Frank are speculators who think Alice's model is commercially promising. Each stakes 2000 DC to increase the upfront guarantee to 10,000 DC.
3. Data scientists build predictive models on the data set and submit them for evaluation.
4. Miners score each model and determine that Bob's model performed the best.
5. Bob provides his model to the DAO in return for his 25% stake and 10,000 DC guarantee. (Alice, Carol, Dan & Eliot own the remaining 75%)
6. The owners make their oil model available for purchase to anyone willing to pay for its execution. Price is set at 10 glu every time the model is executed. (glu is fixed at 100 per DC)
7. The model proves to have strong predictive value and earns over 100,000 DC for its owners. The first 40,000 is shared between Carol, Dan, Eliot and Frank until the up front guarantee is recouped. The subsequent 60,000 includes Bob's share.
8. George runs ShipCo, which provides inventory management solutions for container ships. ShipCo has extensive data on shipment volumes that it thinks will help predict the price of oil.

He conditionally provides his data to understand how much he might get paid.

9. George's data improves predicted performance significantly and the team estimates they can increase the price by 20%. George provides his data in return for a 10% stake in the model.
10. Tired of sharing the model with other investors, Alice offers to purchase other investor's stakes and they accept. Everyone else leaves to invest in other models.

We believe speculators will provide initial incentives to spur model creation. As the commercial potential is revealed, it will attract additional data and more traditional customers/investors. Without the fungibility provided by DC, this fluid transfer is much more difficult.

3 Data Sets

Better training data yields better performing model. Owners of data can supply two types of data to improve a model:

1. More training data
2. Appended data on data records. e.g. A credit score, demographics, or past purchase history may be appended to a borrower's record.

The contribution of each data set is quantified in the same way that the predictive models are evaluated and shares in the DAO are allocated accordingly.

4 Token & Mining

4.1 ERC-20 Implementation

DecisionCoin will be implemented initially as an ERC-20 token. Ethereum provides well-established and tested means for creating DAOs that can handle the trustless revenue share required.

4.2 Dedicated Blockchain

Longer term, there are significant advantages to a dedicated blockchain for Decision Coin. Ethereum's cost for data and processing may prove prohibitive and a mining scheme that supports the token directly is much more attractive. We will provide a timeline for this transition and a discussion of the possible mining schemes in the coming months.

5 Key Entities: Decisions, Models, & Requests

There are several important entities within the DECIDE network: Decisions, Models, & Requests. The following sections provide an outline on those key entities and how they interact. Given the

5.1 Model Requests

Model Requests describe a problem to solve with a predictive model. They outline problem to be solved, and are looking for training a predictive model:

- **Predicted variable:** The target variable that the Solution predicts. e.g. Estimated default probability for a borrower
- **Labeled data set:** The historical data set to be used for training and evaluation. e.g. The historical record of all borrowers with whether they defaulted or not. This is divided into a test data set that is publicly available and a training data set that is used by miners to evaluate the accuracy of the model.
- **Evaluation criteria:** The measure that should be used to determine model accuracy. [Logarithmic loss](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html) is a commonly used metric for classifiers but we should support a cost matrix that will allow the Decision to specifically weight error from false positives, false negatives, et al.

5.2 Data Set Requests

The Data Set version of a Model Request. They outline problem to be solved, and are looking for training data to help solve that problem:

- **Predicted variable:** The target variable that the Solution predicts. e.g. Estimated default probability for a borrower
- **Labeled data set:** The historical data set to be used for training and evaluation. e.g. The historical record of all borrowers with whether they defaulted or not. This is divided into a test data set that is publicly available and a training data set that is used by miners to evaluate the accuracy of the model.
- **Evaluation criteria:** The measure that should be used to determine model accuracy. [Logarithmic loss](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html) is a commonly used metric for classifiers but we should support a cost matrix that will allow the Decision to specifically weight error from false positives, false negatives, et al.

5.3 Models

Models are the proposed solutions to Model Requests. They expose an API that, given one or more labeled items of data, will respond with their estimate of the predicted variable.

- **Decision:** The address of the Decision this solves.
- **Endpoint:** The endpoint where the model lives. (This is hashed so it can only be evaluated by a qualified miner.)
- **Score:** The score when the Model was evaluated against the testing data set.
- **Minimum price:** The minimum amount of DC required to execute the model on a single Request.
- **Payment address:** The address that should receive payment.

5.4 Records

Records are the individual decisions that businesses want to make using Models. Businesses submit Records for evaluation by the best Model within their pricing parameters.

- **Decision:** The address of the Decision.
- **Maximum price:** The maximum amount of DC that the business is willing to pay for having each record evaluated by the model.
- **Number of records:** The number of records the business wishes to evaluate.
- **Expiration time:** How long the Request will remain open.
- **Payment address:** The address that provides payment for the model execution. This will be a proxy that holds Max Price * # of records and will distribute payment after model execution.

5.5 Data Sets

Data Sets consist of the following components:

- **Decision:** The address of the Decision this enriches.
- **Model:** The address of the Model this enriches. Enrichment's impact is model dependent.
- **Endpoint:** The endpoint where the data enrichment lives (for appends)
- **Score improvement:** The increase in score when the data enrichment is used.
- **Minimum price:** The amount of DC required to use the data enrichment on a single Record.

5.6 Creating a bounty

Anyone can create a Bounty by staking DC and specifying what they want to buy:

- **Decision:** The address of the Decision the Model will solve.
- **Bounty:** The DC payable as a bounty for a qualified model.
- **Minimum Score:** The minimum score that needs to be achieved by the model. If none, DECIDE will select the highest scoring Model submitted.
- **First Payable:** When the bounty is first eligible for payment. If none, it's payable to the first model that qualifies.
- **Expiration time:** How long the Bounty will remain open.
- **Payment address:** The address providing payment for the model execution. This will be a proxy that holds the bounty and will distribute payment after the code is transferred and verified.
- **Stakable:** Whether other owners can contribute to the bounty and share in Model ownership.
- **Exclusivity:** Whether the solution is available as a public API or is reserved for the owners.

The DECIDE network will determine when parameters are met and provide a Proxy for that transfer. Initially, verification process will be handled via a manual escrow process but will transition to a decentralized process.

6 Matchmaking & Model Evaluation

During each epoch, miners will do the following:

1. Identify any expired Requests and eliminate them.
2. Identify all matching Requests and Models, execute the models, and transfer DC appropriately.
3. Test any new Models that have been submitted for existing Decisions. (More detail on this process to come, including staking & bounty mechanisms)