Venture capital (VC) is financial capital provided to early-stage, high-potential, growth startup firms. VC firms are the catalysts behind many successful high-tech firms, such as Google, Apple, Microsoft and Alibaba. Recently technological advances have made startups easier to launch than ever before, but they remain competitive and risky endeavours. For typically cashflow-poor startups, financial capital is a key bottleneck for development so entrepreneurs seek VC investment to support them. A VC firm funds a startup company with cash in exchange for an equity stake. It may appear that the dynamics of the transaction are like that of an investor buying shares of a publicly traded company but this is misleading. Rather, venture capitalists have two primary roles: as scouts, able to identify the potential of new startups, and as coaches, able to help startups realise that potential \cite{baum2004}. When a VC funds a startup, the VC often takes an active role in managing the startup, providing expertise and advice in both managerial and technical areas. In this way, the experience and wisdom of the VC’s who invest in a startup directly influence the startup’s trajectory.

VC firms have difficulty finding investments that can provide highly profitable liquidity events. Venture capitalists face the challenge of choosing a few outstanding investments from a sea of thousands of potential opportunities. VC firms seek investments in companies that can provide a profitable liquidity event within the time frame of their fund. For startups, a liquidity event (also commonly referred to as an `exit’) is either an Initial Public Offering (IPO) or an acquisition by a larger competitor. Most VC firms expect their investments to reach a liquidity event within 3--8 years. When compared to traditional investors, this would be considered a long-term investment strategy. However, when compared to the trajectory of most companies, this period is particularly short -- not many companies, even successful companies, are capable of maturing at this rate. This makes the VC investment screening process particularly difficult. In addition, there are many potential investment candidates, most of which ultimately fail, and traditional metrics of performance (e.g. cashflow, earnings) often do not exist or are uncertain \cite{shane2002}. Traditionally, investment opportunities are either referred or identified through technology scans (e.g. Google searches, patent searches). These manual search processes are time-consuming for VC firms and may lead to availability bias.

The VC industry requires better systems and processes to efficiently manage labour-intensive tasks like investment screening. Existing approaches to improve startup investment screening have three common limitations: small sample size, a focus on early-stage investment, and incomplete use of features. The popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for better, more quantitative methods of assessing startup potential. By 2014, over 1,200 investment organisations (including 624 VC firms) were members of CrunchBase's Venture Program, mining CrunchBase's startup data to help inform their investment decisions (CITATION). There is preliminary evidence from the literature that mining these sources may be able to address previous limitations and make investment screening more efficient and effective. Recommendation techniques have the possibility of helping VC firms making data-driven investment decisions by providing an automatic screening process of many startups across different domains based on information about their fundamental characteristics, current performance, and third party validation.

We believe it is now possible to address previous limitations in this domain and produce a VC investment screening system that is efficient, robust and powerful. Our system is based around identifying startup companies that are likely to have a liquidity event (exit) in a given forecast window. This system can generate statistics and make recommendations that may assist venture capitalists to efficiently screen investment candidates. To be useful in this context, the implementation of the system must meet the following criteria:

1. Efficiency

Our system must be more efficient than traditional, manual investment screening. A technique to achieve this is autonomously collecting data from readily-available, online data sources. In this project, we will focus primarily on the CrunchBase online database. We will test whether this source provides a comprehensive feature set and enough observations to provide meaningful statistics. We will also test whether our system can perform its analyses in a reasonable time-period (i.e. at most, a few days) so the results can be used in business decisions.

1. Robustness

Our system must be robust enough to be reliable over time and agnostic to specific data sources. The system must provide a generalised, robust solution for investors that does not require significant technical knowledge to use, is not specific to a time-period, and is not reliant on a single data source. The parameters and features that the system selects should be invariant to time so investors can have reasonable confidence in its predictions. The feature set should be generalisable enough to be data source agnostic and easily interpretable, allowing investors to complement or replace their data sources over time.

1. Predictive Power

Our system must be consistently accurate at identifying a variety of high-potential investment candidates. The system will provide a broad first screening process for investors so it is important that it is highly sensitive. The system should be robust to different forecast windows (i.e. exit in three years from now) as venture capitalists make investment decisions with different periods so they can strategically manage the investment horizons of their funds. Similarly, the system should be able to accurately access companies at any developmental stage.

The following work is presented in three chapters:

1. Literature Review:

We review the theoretical background of startup performance and VC investment and evaluate previous attempts at using data mining in this domain. We determine that the venture capital industry requires better systems to efficiently manage labour-intensive tasks like investment screening. Existing approaches in the literature to tackle similar business problems have three common limitations: small sample size, a focus on early-stage investment, and incomplete use of features. Preliminary evidence suggests that online data sources and machine learning techniques may allow us to address previous limitations and produce an investment screening system that is efficient, robust and powerful.

1. Design & Implementation

We outline our system architecture: covering data collection, pre-processing, classification and our experimental setup. We use data from the CrunchBase online database, with some supplementation from PatentsView (US Patents Office). We use two datasets collected in September 2016 and April 2017 for training and testing respectively. We develop a machine learning pipeline using the popular Python-based machine learning library Scikit-learn (CITATION). Our pipeline includes imputation, variance thresholding, feature transformation, extraction and classification with hyperparameter search optimisation. Our experimental setup also involves reverse-engineering historical datasets using timestamped metadata to test our models’ robustness over time and with respect to our desired forecast windows.

1. Evaluation

We report the results of our experiments on our three criteria: efficiency, robustness and predictive power. Firstly, we evaluate efficiency by exploring the learning curves of our classification techniques and whether there is sufficient data to produce reliable statistics. We also evaluate efficiency by exploring the time profile of our system. Secondly, we evaluate robustness by evaluating our models against multiple reverse-engineered historical datasets and measuring their variance. We also evaluate robustness with respect to accuracy at different levels of feature abstraction (e.g. feature grouping based on our conceptual framework and principal components analysis). Thirdly, we evaluate predictive power by testing different forecast windows outcomes and evaluating our models’ accuracy for startups at different stages of their development lifecycle. Finally, we discuss our findings more broadly and their implications for investors and future research into startup investment and performance.